

# Recovering Structured Probability Matrices

Qingqing Huang<sup>\*</sup>   Sham M. Kakade<sup>†</sup>   Weihao Kong<sup>‡</sup>   Gregory Valiant<sup>§</sup>

## Abstract

We consider the problem of accurately recovering a matrix  $\mathbb{B}$  of size  $M \times M$ , which represents a probability distribution over  $M^2$  outcomes, given access to an observed matrix of “counts” generated by taking independent samples from the distribution  $\mathbb{B}$ . How can structural properties of the underlying matrix  $\mathbb{B}$  be leveraged to yield computationally efficient and information theoretically optimal reconstruction algorithms? When can accurate reconstruction be accomplished in the sparse data regime? This basic problem lies at the core of a number of questions that are currently being considered by different communities, including building recommendation systems and collaborative filtering in the sparse data regime, community detection in sparse random graphs, learning structured models such as topic models or hidden Markov models, and the efforts from the natural language processing community to compute “word embeddings”. Many aspects of this problem—both in terms of learning and property testing/estimation and on both the algorithmic and information theoretic sides—remain open.

Our results apply to the setting where  $\mathbb{B}$  has a low rank structure. For this setting, we propose an efficient (and practically viable) algorithm that accurately recovers the underlying  $M \times M$  matrix using  $\Theta(M)$  samples (where we assume the rank is a constant). This result easily translates to  $\Theta(M)$  sample algorithms for learning topic models with two topics over dictionaries of size  $M$ , and learning hidden Markov Models with two hidden states and observation distributions supported on  $M$  elements. These linear sample complexities are optimal, up to constant factors, in an extremely strong sense: even testing basic properties of the underlying matrix (such as whether it has rank 1 or 2) requires  $\Omega(M)$  samples. We provide an even stronger lower bound where distinguishing whether a sequence of observations were drawn from the uniform distribution over  $M$  observations versus being generated by an HMM with two hidden states requires  $\Omega(M)$  observations. This precludes sublinear-sample hypothesis tests for basic properties, such as identity or uniformity, as well as sublinear sample estimators for quantities such as the entropy rate of HMMs.

---

<sup>\*</sup>MIT. Email: qqh@mit.edu.

<sup>†</sup>University of Washington. Email: sham@cs.washington.edu

<sup>‡</sup>Stanford University. Email: kweihao@gmail.com

<sup>§</sup>Stanford University. Email: valiant@stanford.edu. Gregory and Weihao’s contributions were supported by NSF CAREER Award CCF-1351108, and a research grant from the Okawa Foundation.

# 1 Introduction

Consider an unknown  $M \times M$  matrix of probabilities  $\mathbb{B}$ , satisfying  $\sum_{i,j} \mathbb{B}_{i,j} = 1$ . Suppose one is given  $N$  independently drawn  $(i, j)$ -pairs, sampled according to the distribution defined by  $\mathbb{B}$ . How many draws are necessary to accurately recover  $\mathbb{B}$ ? What can one infer about the underlying matrix based on these samples? How can one accurately test whether the underlying matrix possesses certain properties of interest? How do structural assumptions on  $\mathbb{B}$  — for example, the assumption that  $\mathbb{B}$  has low rank — affect the information theoretic or computational complexity of these questions? For the majority of these tasks, we currently lack both a basic understanding of the computational and information theoretic lay of the land, as well as algorithms that seem capable of achieving the information theoretic or computational limits.

This general question of making accurate inferences about a matrix of probabilities, given a matrix of observed “counts” of discrete outcomes, lies at the core of a number of problems that disparate communities have been tackling independently. On the theoretical side, these problems include both work on community detection in stochastic block models (where the goal is to infer the community memberships from an adjacency matrix of a graph that has been drawn according to an underlying matrix of probabilities expressing the community structure) as well as the line of work on recovering topic models, hidden Markov models (HMMs), and richer structured probabilistic models (where the model parameters can often be recovered using observed count data). On the practical side, these problems include work on computing low-rank approximations to sparsely sampled data, which arise in collaborative filtering and recommendation systems, as well as the recent work from the natural language processing community on understanding matrices of word co-occurrence counts for the purpose of constructing good “word embeddings”. Additionally, work on latent semantic analysis and non-negative matrix factorization can also be recast in this setting.

In this work, we start this line of inquiry by focusing on the estimation problem where the probability matrix  $\mathbb{B}$  possesses a particular low rank structure. While this estimation problem is rather specific, it generalizes the basic community detection problem and also encompasses the underlying problem behind learning HMMs and topic models. Furthermore, this low rank case also provides a means to study how property testing and estimation problems are different in this structured setting, as opposed to the simpler rank 1 setting that is equivalent to the standard setting of independent draws from a distribution supported on  $M$  elements.

We focus on the estimation of a low rank probability matrix  $\mathbb{B}$  in the sparse data regime, near the information theoretic limit. In many practical scenarios involving sample counts, we seek algorithms capable of extracting the underlying structure in the sparsely sampled regime. To give two motivating examples, consider forming the matrix of word co-occurrences—the matrix whose rows and columns are indexed by the set of words, and whose  $(i, j)$ -th element consists of the number of times the  $i$ -th word follows the  $j$ -th word in a large corpus of text. In the context of recommendation system, one could consider a low rank matrix model, where the rows are indexed by customers, and the columns are indexed by products, with the  $(i, j)$ -th entry corresponding to the number of times the  $i$ -th customer has purchased the  $j$ -th product. In both settings, the structure of the probability matrix underlying these observed counts contains insights into the two domains, and in both domains we only have relatively sparse data. This is inherent in many other natural scenarios involving heavy-tailed distributions where, regardless of how much data one collects, a significant fraction of items (e.g. words, products purchased, genetic mutations, etc.) will only be observed a few times.

Such estimation questions have been actively studied in the community detection literature, where the objective is to accurately recover the communities in the regime where the average degree (e.g. the row sums of the adjacency matrix) are constant. In contrast, the recent line of works for recovering highly structured models (such as topic models, HMMs, etc.) are only applicable to the *over-sampled* regime where the amount of data is well beyond the information theoretic limits. In these cases,

achieving the information theoretic limits remains a widely open question. This work begins to bridge the divide between these recent algorithmic advances in both communities. We hope that the low rank probability matrix setting that studied here serves as a jumping-off point for the more general questions of developing information theoretically optimal algorithms for estimating structured matrices and tensors in general, or recovering low-rank approximations to arbitrary probability matrices, in the sparse data regime. While the general settings are more challenging, we believe that some of our algorithmic techniques can be fruitfully extended.

In addition to developing algorithmic tools which we hope are applicable to a wider class of problems, a second motivation for considering this particular low rank case is that, with respect to distribution learning and property testing, the entire lay-of-the-land seems to change completely when the probability matrix  $\mathbb{B}$  has rank larger than 1. In the rank 1 setting — where a sample consists of 2 *independent* draws from a distribution supported on  $\{1, \dots, M\}$  — the distribution can be learned using  $\Theta(M)$  draws. Nevertheless, many properties of interest can be tested or estimated using a sample size that is *sublinear* in  $M^1$ . However, even just in the case where the probability matrix is of rank 2, although the underlying matrix  $\mathbb{B}$  can be represented with  $O(M)$  parameters (and, as we show, it can also be accurately and efficiently recovered with  $O(M)$  sample counts), sublinear sample property testing and estimation is generally impossible. This result begs a more general question: *what conditions must be true of a structured statistical setting in order for property testing to be easier than learning?*

## 1.1 Problem Formulation

Assume our vocabulary is the index set  $\mathcal{M} = \{1, \dots, M\}$  of  $M$  words and that there is an underlying low rank probability matrix  $\mathbb{B}$ , of size  $M \times M$ , with the following structure:

$$\mathbb{B} = PWP^\top, \text{ where matrix } P = [p^{(1)}, \dots, p^{(R)}]. \quad (1)$$

Here the matrix  $P$  is of dimension  $M \times R$ , and the columns are supported on the standard  $(M-1)$ -simplex. Also,  $W \in \mathbb{R}_+^{R \times R}$  is the *mixing matrix*, which is a probability matrix satisfying  $\sum_{i,j} W_{i,j} = 1$ .

In the case where  $R = 2$ , we denote  $w_p = W_{1,1} + W_{1,2}$  and  $w_q = W_{2,1} + W_{2,2}$ . Note that  $\sum_k \mathbb{B}_{i,k} = w_p p + w_q q$ . Define the *covariance matrix* of any probability matrix  $P$  as:

$$[\text{Cov}(P)]_{i,j} := P_{i,j} - \left(\sum_k P_{i,k}\right)\left(\sum_k P_{k,j}\right).$$

Note that  $\text{Cov}(P)\vec{1} = \vec{0}$  and  $\vec{1}^\top \text{Cov}(P) = \vec{0}$  (where  $\vec{1}$  and  $\vec{0}$  are the all ones and zeros vectors, respectively). This implies that, without loss of generality, the covariance of the mixing matrix,  $\text{Cov}(W)$ , can be expressed as:  $\text{Cov}(W) = [w_L, -w_L]^\top [w_R, -w_R]$ , for some real numbers  $w_L, w_R \in [-1, 1]$ . For ease of exposition, we restrict to the symmetric case where  $w_L = w_R = w$ , though our results hold more generally.

Suppose we obtain  $N$ , i.i.d. sample counts from  $\mathbb{B}$  of the form  $\{(i_1, j_1) (i_2, j_2), \dots (i_N, j_N)\}$ , where each sample  $(i_n, j_n) \in \mathcal{M} \times \mathcal{M}$ . The probability of obtaining a count  $(i, j)$  in a sample is  $\mathbb{B}_{i,j}$ . Moreover, assume that the number of samples follows a Poisson distribution:  $N \sim \text{Poi}(\mathbb{N})$ . The Poisson assumption on the number of samples is made only for the convenience of analysis: so that the counts of observing  $(i, j)$  follows a Poisson distribution  $\text{Poi}(\mathbb{N}\mathbb{B}_{i,j})$  and is independent from the counts of observing  $(i', j')$  for  $(i', j') \neq (i, j)$ . As  $M$  is asymptotically large, with high probability,

---

<sup>1</sup>Distinguishing whether a distribution is uniform versus far from uniform can be accomplished using only  $O(\sqrt{M})$  draws, testing whether two sets of samples were drawn from similar distributions can be done with  $O(M^{2/3})$  draws, estimating the entropy of the distribution to within an additive  $\epsilon$  can be done with  $O(\frac{M}{\epsilon \log M})$  draws, etc.

$N$  and  $\mathbb{N}$  are within a subconstant factor of each other and both upper and lower bounds translate between the Poissonized setting, and the setting of fixed  $N$ . Throughout, our sample complexity results are stated in terms of  $N$ .

**Notation** Throughout the paper, we use the following standard shorthand notations.

Denote  $[n] \triangleq \{1, \dots, n\}$ . Let  $\mathcal{I}$  denote a subset of indices in  $\mathcal{M}$ . For a  $M$ -dimensional vector  $x$ , we use vector  $x_{\mathcal{I}}$  to denote the elements of  $x$  restricting to the indices in  $\mathcal{I}$ ; for two index sets  $\mathcal{I}, \mathcal{J}$ , and a  $M \times M$  dimensional matrix  $X$ , we use  $X_{\mathcal{I} \times \mathcal{J}}$  denote the submatrix of  $X$  with rows restricting to indices in  $\mathcal{I}$  and columns restricting to indices in  $\mathcal{J}$ .

We use  $\text{Poi}(\lambda)$  to denote a Poisson distribution with rate  $\lambda$ ; we use  $\text{Ber}(\lambda)$  to denote a Bernoulli random variable with success probability  $\lambda$ ; and we use  $\text{Mul}(x; \lambda)$  to denote a multinomial distribution over  $M$  outcomes with  $\lambda$  number of trials and event probability vector  $x \in \mathbb{R}_+^M$  such that  $\sum_i x_i = 1$ .

All of our order notations are with respect to the vocabulary size  $M$ , which is asymptotically large. Also, we say that a statement is true “with high probability” if the failure probability of the statement is inverse poly in  $M$ ; and we say a statement is true “with large probability” if the failure probability is of some small constant  $\delta$ , which can be easily boosted via repetition.

## 1.2 Main Results

### 1.2.1 Recovering Low Rank Probability Matrices

For rank  $R = 2$ , it is possible to recover the dictionary  $P = [p, q]$  uniquely up to column permutation. Assume that  $W$  is symmetric, where  $w_L = w_R = w$  (all our results extend to the asymmetric case). Define the *marginal probability* vector,  $\rho$  and the *dictionary separation* vector as:

$$\rho_i := \sum_k \mathbb{B}_{i,k}, \quad \Delta := w(p - q). \quad (2)$$

Observe that in this rank 2 case, the matrix  $\text{Cov}(\mathbb{B})$  admits a unique rank-1 decomposition, which implies that:

$$\mathbb{B} = \rho \rho^\top + \Delta \Delta^\top. \quad (3)$$

We focus on a class of model parameters where  $p$  and  $q$  are well separated, which assumption guarantees that the rank 2 matrix  $\mathbb{B}$  is well-conditioned. This assumption also has natural interpretations in different applications including community detection, topic modeling, and HMMs.

**Assumption 1** (Separation). *Assume that  $w_p$  and  $w_q$  are lower bounded by some constant  $C_w = \Omega(1)$ , and assume that the  $\ell_1$ -norm of the dictionary separation is lower bounded by  $\|\Delta\|_1 \geq C_\Delta = \Omega(1)$ .*

**Theorem 1.1** (Upper bound for rank 2 matrices). *Suppose we have access to  $N$  i.i.d. samples generated according to the a rank 2 symmetric probability matrix  $\mathbb{B}$  parameterized as (1), and suppose the true matrix satisfies Assumption 1. For  $\epsilon > 0$ , with  $N = \Theta(M/\epsilon^2)$  samples, our algorithm runs in time  $\text{poly}(M)$  and returns estimators  $\hat{B}, \hat{\rho}, \hat{\Delta}$ , such that with large probability:*

$$\|\hat{B} - \mathbb{B}\|_1 \leq \epsilon, \quad \|\hat{\rho} - \rho\|_1 \leq \epsilon, \quad \|\hat{\Delta} - \Delta\|_1 \leq \epsilon.$$

(here, the  $\ell_1$ -norm of an  $M \times M$  matrix  $P$  is simply defined as  $\|P\|_1 = \sum_{i,j} |P_{i,j}|$ ).

Note that for  $R > 2$ , the dictionary matrix  $P$  and the mixing matrix  $W$  are not uniquely identifiable. We only focus on obtaining a low rank estimator for the underlying probability matrix  $\mathbb{B}$ .

**Theorem 1.2** (Upper bound for rank  $R$ , constant accuracy). *Suppose we have access to  $N$  i.i.d. samples generated according to the a probability matrix  $\mathbb{B}$  parameterized as (1). Assume the mixing matrix  $W$  is PSD with row sums bounded by  $\sum_j W_{i,j} \geq w_{\min}$ . Fix constant accuracy  $\epsilon_0 > 0$  and  $\epsilon_0 = \Omega(1)$ , for any  $r > 0$ , with  $N = \Theta(\frac{MR^2}{w_{\min}^2 \epsilon_0^{4+r}})$  samples, our algorithm runs in time  $\text{poly}(M)$  and returns a rank  $R$  estimator  $\hat{B}$  such that with large probability:*

$$\|\hat{B} - \mathbb{B}\|_1 \leq \epsilon_0. \quad (4)$$

Compared to the sample complexity result  $N = \Theta(MR^2)$  for the community detection problem with  $R$  communities as in [19], in the more general parameterization, we incur an extra  $w_{\min}^{-2}$  dependence, which can be easily removed in the special setup of community detection to recover the result in the community detection problem.

**Assumption 2** (Well separated dictionary). *We assume that the minimal singular value of  $\mathbb{B}^{sqr}$  scaled with the inverse square root of the exact marginal probabilities is lower bounded.*

$$\sigma_R(\text{Diag}(\rho_i)^{-1/2} \mathbb{B} \text{Diag}(\rho_i)^{-1/2}) \geq \sigma_{\min}. \quad (5)$$

Note that in the ideal case where the support of the dictionaries are non-overlapping, and the mixing matrix  $W$  is diagonal, we have  $\sigma_1(\text{Diag}(\rho_i)^{-1/2} \mathbb{B} \text{Diag}(\rho_i)^{-1/2}) = \sigma_R(\text{Diag}(\rho_i)^{-1/2} \mathbb{B} \text{Diag}(\rho_i)^{-1/2}) = 1$ .

Under the well-separation assumption for the dictionary, we can sharpen the error bound.

**Theorem 1.3.** (Upper bound for rank  $R$  under separation condition) *Under the conditions of Theorem 1.2, further assume that Assumption 2 is satisfied for  $\sigma_{\min} > \epsilon_0$ , and that  $N = \Omega(\frac{MR^2}{w_{\min}^2 \epsilon_0^{4+r}})$  for any  $r > 0$ . For any  $\epsilon > 0$  such that  $\epsilon < \epsilon_0$ , with  $N = \Theta(\frac{MR}{\epsilon^2})$  samples, our algorithm runs in time  $\text{poly}(M)$  and returns a rank  $R$  estimator  $\hat{B}$  such that with large probability:*

$$\|\hat{B} - \mathbb{B}\|_1 \leq \epsilon. \quad (6)$$

Note that when the marginal probabilities  $\rho_i$  are not roughly uniform, spectral error bounds in terms of  $\|\hat{B} - \mathbb{B}\|_2$  are not particularly strong. Instead, here we consider the  $\ell_1$  norm error bound, or equivalently the total variation distance, which is a more natural measure of estimation error for probability distributions. Moreover, note that naively estimating a distribution over  $M^2$  outcomes requires order  $M^2$  samples. Our algorithm utilizes the low rank structure of the underlying probability matrix to achieve a sample complexity which is *precisely* linear in the vocabulary size  $M$ .

We now turn to the implications of this theorem to testing and learning problems.

### 1.2.2 Topic Models and Hidden Markov Models

One of the main motivations for considering the specific low rank structure on the underlying matrix  $\mathbb{B}$  is that this structure encompasses the structure of the matrix of expected bigrams generated by both topic models and HMMs. We now make these connections explicit for the rank 2 case, and then briefly discuss the rank  $R$  case.

**Definition 1.4.** *A 2-topic model over a vocabulary of size  $M$  is defined by a pair of distributions,  $p$  and  $q$  supported over  $M$  words, and a pair of topic mixing weights  $\pi_p$  and  $\pi_q = 1 - \pi_p$ . The process of drawing a bigram  $(i, j)$  consists of first randomly picking one of the two “topics” according to the mixing weights, and then drawing two independent words from the word distribution corresponding to the chosen topic. Thus the probability of seeing bigram  $(i, j)$  is  $(\pi_p p_i p_j + \pi_q q_i q_j)$ , and so the expected bigram matrix can be written as  $\mathbb{B} = PWP^\top$  with  $P = [p, q]$ , and  $W = [\pi_p, 0; 0, \pi_q]$ .*

**Definition 1.5.** A hidden Markov model with 2 hidden states  $(s_p, s_q)$  and a size  $M$  observation vocabulary is defined by a  $2 \times 2$  transition matrix  $T$  for the 2 hidden states, and two distributions of observations,  $p$  and  $q$ , corresponding to the 2 states.

A sequence of  $N$  observations is sampled as follows: First, select an initial state according to the stationary distribution of the underlying Markov chain  $[\pi_p, \pi_q]$ ; Then evolve the Markov chain according to the transition matrix  $T$  for  $N$  steps; For each  $n \in \{1, \dots, N\}$ , the  $n$ -th observation in the sequence is generated by making an independent draw from either distribution  $p$  or  $q$  according to whether the Markov chain is in state  $s_p$  or  $s_q$  at the  $n$ -th timestep.

The probability that seeing a bigram  $(i, j)$  for the  $n$  and the  $(n + 1)$ -th observation is given by  $\pi_p p_i (T_{p,p} p_j + T_{p,q} q_j) + \pi_q q_i (T_{q,p} p_j + T_{q,q} q_j)$ , and hence the expected bigram matrix can be written as  $\mathbb{B} = D W D^\top$  with  $D = [p, q]$ , and  $W = \begin{bmatrix} \pi_p & 0 \\ 0 & \pi_q \end{bmatrix} \begin{bmatrix} T_{p,p} & 1 - T_{p,p} \\ 1 - T_{q,q} & T_{q,q} \end{bmatrix}$ .

The following corollaries (straightforward by Theorem 1.1) shows that parameter estimation is possible with sample size *linear* in  $M$ :

**Corollary 1.6.** (*Learning 2-topic models*) Suppose we are in the 2-topic model setting. Assume that  $\pi_p(1 - \pi_p)\|p - q\|_1 = \Omega(1)$ . There exists an algorithm which, given  $N = \Omega(M/\epsilon^2)$  bigrams, runs in time  $\text{poly}(M)$  and with large probability returns estimates  $\hat{\pi}_p, \hat{p}, \hat{q}$  such that

$$|\hat{\pi}_p - \pi_p| < \epsilon, \|\hat{p} - p\|_1 \leq \epsilon, \|\hat{q} - q\|_1 \leq \epsilon.$$

**Corollary 1.7.** (*Learning 2-state HMMs*) Suppose we are in the 2-state HMM setting. Assume that  $\|p - q\|_1 \geq C_1$  and that  $\pi_p, T_{p,p}, T_{q,q}$  are lower bounded by  $C_2$  and upper bounded by  $1 - C_2$ , where both  $C_1$  and  $C_2$  are  $\Omega(1)$ . There exists an algorithm which, given a sampled chain of length  $N = \Omega(M/\epsilon^2)$ , runs in time  $\text{poly}(M)$  and returns estimates  $\hat{\pi}_p, \hat{T}, \hat{p}, \hat{q}$  such that, with high probability, we have (that there is exists a permutation of the model such that)

$$|\hat{\pi}_p - \pi_p| < \epsilon, |\hat{T}_{p,p} - T_{p,p}| < \epsilon, |\hat{T}_{q,q} - T_{q,q}| < \epsilon, \|\hat{p} - p\|_1 \leq \epsilon, \|\hat{q} - q\|_1 \leq \epsilon.$$

Furthermore, it is sufficient for this algorithm to only utilize  $\Omega(M/\epsilon^2)$  random bigrams and only  $\Omega(1/\epsilon^2)$  random trigrams from this chain.

For topic models with  $R > 2$  topics and HMMs with  $R > 2$  hidden states, the matrix of bigram probabilities does not uniquely determine the underlying HMM. One can recover the model parameters using sampled trigram sequences (see [7] for the moment structure in the trigrams). However, the core step remains to first obtain an accurate estimate of  $\mathbb{B}$  given by Theorem 1.2 and 1.3<sup>2</sup>. We do not go into details in this draft.

### 1.2.3 Testing vs. Learning

The above theorem and corollaries are tight in an extremely strong sense: for both the topic model and HMM settings, it is information theoretically impossible to perform even the most basic property tests using fewer than  $\Theta(M)$  samples. For topic models, the community detection lower bounds [41][32][52] imply that  $\Theta(M)$  bigrams are necessary to even distinguish between the case that the underlying model is simply the uniform distribution over bigrams versus the case of a  $R$ -topic model in which each topic corresponds to a uniform distribution over disjoint subsets of  $M/R$  words. For 2-state HMMs, even if we permit an estimator to have more information than merely bigram counts, namely the *full sequence* of observations, we prove the following linear lower bound.

<sup>2</sup>E.g. see [7] for how the bigram matrix can be used in the estimation problem in a “whitening” step to reduce the problem from one of  $M$  dimensions to one with effectively  $R$  dimensions.

**Theorem 1.8.** *There exists a constant  $c > 0$  such that for sufficiently large  $M$ , given a sequence of observations from a HMM with two states and emission distributions  $p, q$  supported on  $M$  elements, even if the underlying Markov process is symmetric, with transition probability  $1/4$ , it is information theoretically impossible to distinguish the case that the two emission distributions,  $p = q = \text{Unif}[M]$  from the case that  $\|p - q\|_1 = 1$  with probability greater than  $2/3$  using a sequence of fewer than  $cM$  observations.*

This immediately implies the following corollary for estimating the *entropy rate* of an HMM.

**Corollary 1.9.** *There exists an absolute constant  $c > 0$  such that given a sequence of observations from a HMM with two hidden states and emission distributions supported on  $M$  elements, a sequence of  $cM$  observations is information theoretically necessary to estimate the entropy rate to within an additive  $0.5$  with probability of success greater than  $2/3$ .*

These strong lower bounds for property testing and estimation are striking for several reasons. First, the core of our learning algorithm is a matrix reconstruction step that uses only the set of bigram counts. Conceivably, one could significantly benefit from considering longer sequences of observations — even for HMMs that mix in constant time, there are detectable correlations between observations separated by  $O(\log M)$  steps. Regardless, our lower bound shows that actually no additional information from such longer  $k$ -grams can be leveraged to yield sublinear sample property testing or estimation.

A second notable point is the apparent brittleness of sublinear property testing and estimation as we deviate from the standard (unstructured) i.i.d sampling setting. Indeed for nearly all distributional property estimation or testing tasks, including testing uniformity and estimating the entropy, sublinear-sample testing and estimation is possible in the i.i.d. sampling setting (e.g. [25, 49, 48]). In contrast to the i.i.d. setting in which estimation and testing require asymptotically fewer samples than *learning*, as the above results illustrate, even in the setting of an HMM with just two hidden states, learning and testing require comparable numbers of observations.

### 1.3 Related Work

As mentioned earlier, the general problem of reconstructing an underlying matrix of probabilities given access to a count matrix drawn according to the corresponding distribution, lies at the core of questions that are being actively pursued by several different communities. We briefly describe these questions, and their relation to the present work.

**Community Detection.** With the increasing prevalence of large scale social networks, there has been a flurry of activity from the algorithms and probability communities to both model structured random graphs, and understand how (and when it is possible) to examine a graph and infer the underlying structures that might have given rise to the observed graph. One of the most well studied community models is the *stochastic block model* [27]. In its most basic form, this model is parameterized by a number of individuals,  $M$ , and two probabilities,  $\alpha, \beta$ . The model posits that the  $M$  individuals are divided into two equal-sized “communities”, and such a partition defines the following random graph model: for each pair of individuals in the same community, the edge between them is present with probability  $\alpha$  (independently of all other edges); for a pair of individuals in different communities, the edge between them is present with probability  $\beta < \alpha$ . Phrased in the notation of our setting, the adjacency matrix of the graph is generated by including each potential edge  $(i, j)$  independently, with probability  $\mathbb{B}_{i,j}$ , with  $\mathbb{B}_{i,j} = \alpha$  or  $\beta$  according to whether  $i$  and  $j$  are in the same community. Note that  $\mathbb{B}$  has rank 2 and is expressible in the form of Equation 1 as  $\mathbb{B} = PWP^\top$  where  $P = [p, q]$  for vectors  $p = \frac{2}{M}I_1$  and  $q = \frac{2}{M}I_2$  where  $I_1$  is the indicator vector for membership in the first community, and  $I_2$  is defined analogously, and  $W$  is the  $2 \times 2$  matrix with  $\alpha \frac{M^2}{4}$  on the diagonal and  $\beta \frac{M^2}{4}$  on the off-diagonal.

What values of  $\alpha, \beta$ , and  $M$  enable the community affiliations of all individuals to be accurately recovered with high probability? What values of  $\alpha, \beta$ , and  $M$  allow for the graph to be distinguished from an Erdos-Renyi random graph (that has no community structure)? The crucial regime is where  $\alpha, \beta = O(\frac{1}{M})$ , and hence each person has a constant, or logarithmic expected degree. The naive spectral approaches will fail in this regime, as there will likely be at least one node with degree  $\approx \log M / \log \log M$ , which will ruin the top eigenvector. Nevertheless, in a sequence of works sparked by the paper of Friedman, and Szemerédi [22], the following punchline has emerged: the naive spectral approach will work, even in the constant expected degree setting, provided one first either removes, or at least diminishes the weight of these high-degree problem vertices (e.g. [21, 31, 40, 32, 33]). In the past year, for both the *exact* recovery problem and the detection problem, the exact tradeoffs between  $\alpha, \beta$ , and  $M$  were established, down to subconstant factors [41, 1, 36]. More recently, there has been further research investigating more complex stochastic block models, consisting of three or more components, components of unequal sizes, etc. (see e.g. [19, 2, 3]).

**Word Embeddings.** On the more applied side, some of the most impactful advances in natural language processing over the past two years has been work on “word embeddings” [37, 35, 46, 10]. The main idea is to map every word  $w$  to a vector  $v_w \in \mathbb{R}^d$  (typically  $d \approx 500$ ) in such a way that the geometry of the vectors captures the semantics of the word.<sup>3</sup> One of the main constructions for such embeddings is to form the  $M \times M$  matrix whose rows/columns are indexed by words, with  $(i, j)$ -th entry corresponding to the total number of times the  $i$ -th and  $j$ -th word occur next to (or near) each other in a large corpus of text (e.g. wikipedia). The word embedding is then computed as the rows of the singular vectors corresponding to the top rank  $d$  approximation to this empirical count matrix.<sup>4</sup> These embeddings have proved to be extremely effective, particularly when used as a way to map text to features that can then be trained in downstream applications. Despite their successes, current embeddings seem to suffer from sampling noise in the count matrix (where many transformations of the count data are employed, e.g. see [45])—this is especially noticeable in the relatively poor quality of the embeddings for relatively rare words. The recent theoretical work [11] sheds some light on why current approaches are so successful, yet the following question largely remains: Is there a more accurate way to recover the best rank- $d$  approximation of the underlying matrix than simply computing the best rank- $d$  approximation for the (noisy) matrix of empirical counts?

**Efficient Algorithms for Latent Variable Models.** There is a growing body of work from the algorithmic side (as opposed to information theoretic) on how to recover the structure underlying various structured statistical settings. This body of work includes work on learning HMMs [29, 39, 18], recovering low-rank structure [9, 8, 15], and learning or clustering various structured distributions such as Gaussian mixture models [20, 51, 38, 14, 28, 30, 23] and latent dirichlet allocation (a very popular topic model) [6]. A number of these methods essentially can be phrased as solving an inverse moments problem, and the work in [7] provides a unifying viewpoint for computationally efficient estimation for many of these models under a tensor decomposition perspective. In general, this body of work has focussed on the computational issues and has considered these questions in the regime in which the amount of data is plentiful—well above the information theoretic limits.

**Sublinear Sample Testing and Estimation.** In contrast to the work described in the previous section on efforts to devise computationally efficient algorithms for tackling complex structural settings in the “over-sampled” regime, there is also significant work establishing information theoretically

---

<sup>3</sup>The goal of word embeddings is not just to cluster similar words, but to have semantic notions encoded in the geometry of the points: the example usually given is that the direction representing the difference between the vectors corresponding to “king” and “queen” should be similar to the difference between the vectors corresponding to “man” and “woman”, or “uncle” and “aunt”, etc.

<sup>4</sup>A number of pre-processing steps have been considered, including taking the element-wise square roots of the entries, or logarithms of the entries, prior to computing the SVD.



optimal algorithms and (matching) lower bounds for estimation and distributional hypothesis testing in the most basic setting of independent samples drawn from (unstructured) distributions. This work includes algorithms for estimating basic statistical properties such as entropy [43, 26, 47, 49], support size [44, 47], distance between distributions [47, 49, 48], and various hypothesis tests, such as whether two distributions are very similar, versus significantly different [25, 12, 42, 50, 16], etc. While many of these results are optimal in a worst-case (“minimax”) sense, there has also been recent progress on instance optimal (or “competitive”) estimation and testing, e.g. [4, 5, 50], with stronger information theoretic optimality guarantees. There has also been a long line of work beginning with [17, 13] on these tasks in “simply structured” settings, e.g. where the domain of the distribution has a total ordering or where the distribution is monotonic or unimodal.

## 2 Outline of our estimation algorithm

Given  $N$  samples drawn according to the probability matrix  $\mathbb{B}$ . Let  $B$  denote the matrix of average empirical counts. By the Poisson assumption on sample size, we have that  $[B]_{i,j} \sim \frac{1}{N} \text{Poi}(N\mathbb{B}_{i,j})$ .

Before introducing our algorithm, let us consider the naive approach of estimating  $\mathbb{B}$  by taking the rank  $R$  truncated SVD of the empirical matrix  $B$ , which concentrates to  $\mathbb{B}$  in spectral distance asymptotically. Unfortunately, this approach leads to a sample complexity as large as  $\Theta(M^2 \log M)$ , and in the linear sample size regime, the empirical counts matrix is a poor representation of the underlying distribution. Intuitively, due to the sampling noise, the rows and columns of  $B$  corresponding to words with larger marginal probabilities have higher row and column sums in expectation, as well as higher variances that undermine the spectral concentration of the matrix as a whole. This observation leads to the idea of pre-scaling the matrix so that every word (i.e. row/column) is roughly of unit variance. Indeed, with a slight modification of the truncated SVD, we can improve the sample complexity of this approach to  $\Theta(M \log M)$ , which is nearly linear. Interestingly, if we get to observe a matrix  $(\mathbb{B} + E)$  where the noise matrix  $E$  are i.i.d. sub-Gaussian variables of unit variance, then truncated SVD indeed gives us the optimal estimator for  $\mathbb{B}$ . Our algorithm shows that we can actually shave off the log factor for a broad class of noise (sub-exponential), which require more careful steps than truncated SVD to denoise the empirical matrix.

Next, we sketch the outline of our algorithms (Algorithm 1 for rank 2 case and Algorithm 2 for general rank  $R$  case). We only highlight the intuition behind the key ideas, and defer the detailed analysis of the algorithms to Section 3 and 4.

### 2.1 Rank 2 algorithm

First, note that it is straightforward to obtain an estimate  $\hat{\rho}$  close to the true marginal  $\rho$  with linear sample complexity. Also, recall that  $\mathbb{B} - \rho\rho^\top = \Delta\Delta^\top$  as per (3), hence after subtracting off the relatively accurate rank 1 matrix of  $\hat{\rho}\hat{\rho}^\top$ , we are essentially left with a rank 1 matrix recovery problem. Our Algorithm 1 consists of two phases:

**Phase I: “binning” and “regularization”** In Section 1, we drew the connection between our problem and the community detection problem in sparse random graphs. Recall that when the word marginals are roughly uniform, namely all in the order of  $O(\frac{1}{M})$ , the linear sample regime corresponds to the stochastic block model setup where the expected row sums are all in the order of  $d_0 = \frac{N}{M} = \Omega(1)$ . It is well-known that in this sparse regime, the adjacency matrix, or the empirical count matrix  $B_N$  in our problem, does not concentrate to the expectation matrix in the spectral distance. Due to some heavy rows with row sum in the order of  $\Omega(\frac{\log M}{\log \log M})$ , the leading eigenvectors are polluted by the local properties of these heavy nodes and do not reveal the global structure of the graph, which are precisely the desired information in expectation.

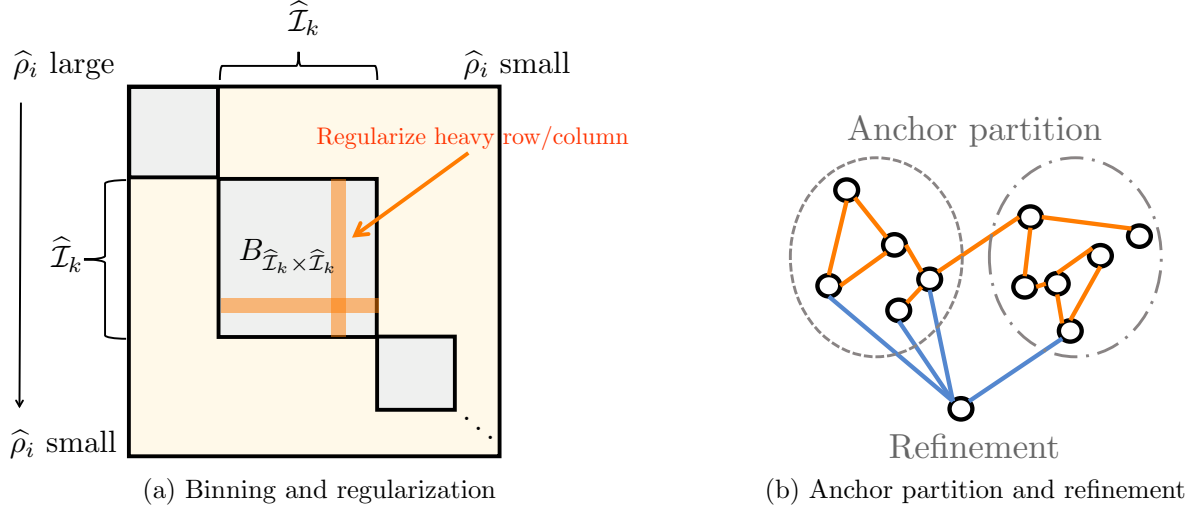


Figure 1: The key algorithmic ideas of our algorithm.

In order to enforce spectral concentration in the linear sample size regime, one of the many techniques is to tame the heavy rows and columns by setting them to 0. This simple idea was first introduced by [22], and followed by analysis works in [21] and many others. Recently in [33] and [34] the authors provided clean and clever proofs to show that *any* such “regularization” essentially leads to better spectral concentration for the adjacency matrix of random graphs whose row/column sums are roughly uniform in expectation.

Phase I of Algorithm 1 leverage such “regularization” ideas in our problem where the marginal probabilities are not uniform with the idea of “binning”. A natural candidate solution would be to partition the vocabulary  $\mathcal{M}$  into bins of words according to the word marginals, so that the words in the same bin have roughly uniform marginals. Restricting our attention to the diagonal blocks of  $\mathbb{B}$  whose indices are in the same bin, the expected row and column sums are indeed roughly uniform. Then we can regularize (by removing abnormally heavy rows and columns) each diagonal block separately to restore spectral concentration, to which truncated SVD should then apply. Figure 1a visualizes the two operations of “binning” and “regularization” in Phase I of Algorithm 1.

Phase I returns estimates  $\hat{\rho}$  and  $\hat{\Delta}$  both up to a small constant accuracy in  $\ell_1$  norm with  $\Theta(M)$  samples. There are 3 concerns we rigorously address in order to prove the correctness of the algorithm:

1. We do not have access to the exact marginal  $\rho$ . With linear sample size, we only can estimate  $\rho$  up to constant accuracy in  $\ell_1$  norm. If we implement binning according to the empirical marginals, there is considerable probability with which words with large marginals are placed in a bin intended for words with small marginals — which we call “spillover effect”. When directly applied to the empirical bins with such spillover, the existing results of “regularization” in [34] do not lead to the desired concentration result.
2. When restricting to each diagonal block corresponding to a bin, we throw away all the sample counts outside the block. This greatly reduces the effective sample size, and it is not obvious that we retain enough samples in each diagonal block to guarantee meaningful estimation.
3. Even if the “regularization” trick works for each diagonal block, we need to extract the useful information and “stitch” together this information from each block to provide an estimator for the entire matrix, including the off-diagonal blocks.

**Phase II: “Anchor partition”** Under the separation Assumption 1, Phase II of Algorithm 1 refine the estimates of Phase I to achieve the desired sample complexity bound.

The key to this refining process is to construct an “anchor partition”, which is a bi-partition of the vocabulary  $\mathcal{M}$  based on the signs of the estimate of separation vector  $\hat{\Delta}$  given by Phase I. We collapse the  $M \times M$  matrix  $B$  into a  $2 \times 2$  matrix corresponding to the bi-partition, and accurately estimate the  $2 \times 2$  matrix with the  $N$  samples. Given this extremely accurate estimate of this  $2 \times 2$  anchor matrix, we can now iteratively refine our estimates of  $\rho_i$  and  $\Delta_i$  for each word  $i$  by solving a simple least square fitting problem.

Similar ideas — estimation refinement based on some crude global information — has appeared in many works for different problems. For example, in a recent paper [19] on community detection, after obtaining a crude classification of nodes using spectral algorithm, one round of a “correction” routine is applied to each node based on its connections to the graph partition given by the first round. This refinement immediately leads to an optimal rate of recovery. Figure 1b visualize the example of community detection. In our problem, the nodes are the  $M$  words, the edges are the sample counts, and instead of re-assigning the label to each node in the refinement routine, and we refine the estimation of  $\rho_i$  and  $\Delta_i$  for each word.

## 2.2 Rank $R$ algorithm

We summarize the basic ideas of Algorithm below. In Step 1, we again group words according to the empirical marginal probabilities, so that in each bin words are of similar marginals. Then in Step 2, we consider the diagonal blocks of the empirical average bigram matrix  $B$ , which rows and columns correspond to the words in the same bin. In each of such diagonal blocks, the entries have roughly uniform expectations, similar to Phase 1 of Algorithm 1, we regularize each diagonal block in the empirical matrix by removing abnormally heavy rows and columns, and then apply truncated SVD to obtain a sharper concentration bound.

After estimate the span of the dictionary restricted to words in each bin by looking at the leading rank  $R$  subspace of each diagonal block, in Step 3, we aim to estimate  $\text{Diag}(\rho)^{1/2} \mathbb{B} \text{Diag}(\rho)^{1/2}$  accurately in spectral norm. With the marginal probability scaling, such error bound naturally translates into error bound for estimating  $\mathbb{B}$  in  $\ell_1$  norm. To achieve this, we regularize and approximately scale the empirical matrix  $B$  with the empirical marginal probability, and then project the entire matrix to a  $R \log M$ -dimensional subspace as a union of the spans for each bin found in Step 2. Since such spans are estimated accurately enough, projecting the  $M \times M$  dimensional matrix to the  $R \log M$ -dimensional subspace preserves the signal that is correlated with the expectation while significantly reducing the statistical noise from sampling. This guarantees a sharp spectral concentration to the expectation  $\text{Diag}(\rho)^{1/2} \mathbb{B} \text{Diag}(\rho)^{1/2}$ .

In the last step, similar to the Phase II of Algorithm 1, if the underlying true probability is “well-conditioned” we can further improve the sample complexity by refine the estimation.

**Input:**  $2N$  sample counts.

**Output:** Estimates  $\hat{\rho}$ ,  $\hat{\Delta}$ ,  $\hat{B}$ .

Divide the sample counts into two independent batches of equal size  $N$ , and construct two average empirical matrices. Each of the following two steps uses an independent copy of  $B$ .

### Phase I

#### 1. Binning according to the empirical marginal probabilities

Set  $\hat{\rho}_i = \sum_{j \in [M]} [B]_{i,j}$ . Partition the vocabulary  $\mathcal{M}$  into:

$$\hat{\mathcal{I}}_0 = \{i : \hat{\rho}_i < \frac{\epsilon_0}{M}\}, \hat{\mathcal{I}}_{\log} = \left\{i : \hat{\rho}_i > \frac{\log(M)}{M}\right\}, \hat{\mathcal{I}}_k = \left\{i : \frac{e^k}{M} \leq \hat{\rho}_i \leq \frac{e^{k+1}}{M}\right\}, k = 1 : \log \log(M).$$

#### 2. Estimate separation vector in each bin (up to sign flip). Set $\hat{\Delta}_{\hat{\mathcal{I}}_0} = 0$ .

If  $\sum \hat{\rho}_{\hat{\mathcal{I}}_{\log}} < \epsilon_0$ , set  $\hat{\Delta}_{\hat{\mathcal{I}}_{\log}} = 0$ , else

**(Rescaling):** Set  $E = \text{diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})^{-1/2} [B - \hat{\rho} \hat{\rho}^\top]_{\hat{\mathcal{I}}_{\log} \times \hat{\mathcal{I}}_{\log}} \text{diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})^{-1/2}$ .

**(SVD):** Let  $u_{\log} u_{\log}^\top$  be the rank-1 truncated SVD of  $E$ . Set  $v_{\log} = \text{diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})^{1/2} u_{\log}$ .

If  $\sum \hat{\rho}_{\hat{\mathcal{I}}_k} < \epsilon_0 e^{-k}$ , set  $\hat{\Delta}_{\hat{\mathcal{I}}_k} = 0$ , else

**(Regularization):** Set  $d_k^{\max} = (\sum \hat{\rho}_{\hat{\mathcal{I}}_k}) \frac{e^{k+\tau}}{M}$ , if a row/column of  $[B]_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k}$  has sum larger than  $2d_k^{\max}$ , set the entire row/column to 0. Let  $\tilde{B}$  denote the regularized block.

**(SVD):** Let  $v_k v_k^\top$  be the rank-1 truncated SVD of  $(\tilde{B} - \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top)$ .

#### 3. Stitching the segments. Fix $k^* = \arg \max_k \|v_k\|$ , set $\hat{\Delta}_{\hat{\mathcal{I}}_{k^*}} = v_{k^*}$ .

For all  $k$ , define  $\mathcal{I}_k^+ = \{i : i \in \hat{\mathcal{I}}_k : \hat{\Delta}_i > 0\}$  and  $\mathcal{I}_k^- = \hat{\mathcal{I}}_k \setminus \mathcal{I}_k^+$ .

Set  $\hat{\Delta}_{\hat{\mathcal{I}}_k} = v_k$  if  $\frac{\sum_{i \in \mathcal{I}_{k^*}^+, j \in \mathcal{I}_k^+} [B]_{i,j}}{\sum_{i \in \mathcal{M}, j \in \mathcal{I}_k^+} [B]_{i,j}} > \frac{\sum_{i \in \mathcal{I}_{k^*}^+, j \in \mathcal{I}_k^-} [B]_{i,j}}{\sum_{i \in \mathcal{M}, j \in \mathcal{I}_k^-} [B]_{i,j}}$ , and  $\hat{\Delta}_{\hat{\mathcal{I}}_k} = -v_k$  otherwise.

### Phase II

#### 1. (Construct anchor partition) Set $\mathcal{A} = \emptyset$ . For all empirical bins, if $\|\hat{\Delta}_{\hat{\mathcal{I}}_k}\|_2 \leq (\sqrt{d_k^{\max}/N})^{1/2}$ , skip the bin; otherwise set $\mathcal{A} = \mathcal{A} \cup \{i \in \hat{\mathcal{I}}_k : \hat{\Delta}_i > 0\}$ .

#### 2. (Estimate anchor matrix)

Set  $B_{\mathcal{A}} = \begin{bmatrix} \sum_{i \in \mathcal{A}, j \in \mathcal{A}} [B_N]_{i,j} & \sum_{i \in \mathcal{A}, j \in \mathcal{A}^c} [B_N]_{i,j} \\ \sum_{i \in \mathcal{A}^c, j \in \mathcal{A}} [B_N]_{i,j} & \sum_{i \in \mathcal{A}^c, j \in \mathcal{A}^c} [B_N]_{i,j} \end{bmatrix}$ . Set vector  $b = \begin{bmatrix} \sum_{i \in \mathcal{A}, j \in \mathcal{M}} [B_N]_{i,j} \\ \sum_{i \in \mathcal{A}^c, j \in \mathcal{M}} [B_N]_{i,j} \end{bmatrix}$ .

Set  $aa^\top$  to be rank-1 truncated SVD of the  $2 \times 2$  matrix  $(B_{\mathcal{A}} - bb^\top)$ .

#### 3. (Refine the estimation:)

$$\text{Set } \begin{bmatrix} \hat{\rho}^\top \\ \hat{\Delta}^\top \end{bmatrix} = [a, b]^{-1} \begin{bmatrix} \sum_{i \in \mathcal{A}} [B_N]_{i, \mathcal{M}} \\ \sum_{i \in \mathcal{A}^c} [B_N]_{i, \mathcal{M}} \end{bmatrix}$$

**Return**  $\hat{\rho}$ ,  $\hat{\Delta}$ , and  $\hat{B} = \hat{\rho} \hat{\rho}^\top + \hat{\Delta} \hat{\Delta}^\top$ .

**Algorithm 1:** Rank 2 algorithm

**Input:**  $4N$  i.i.d. samples from the distribution  $\mathbb{B}$  of dimension  $M \times M$ .

(In each of the 4 steps,  $B$  refers to an independent copy of the bigram matrix with  $N$  samples.)

**Output:** Rank  $R$  estimator  $\hat{B}$  for  $\mathbb{B}$ , and  $\hat{V}$  for the rank  $R$  subspace of scaled matrix  $D_S \mathbb{B}^{sqr t}$ .

Step 1. (**Binning according to the empirical marginal probabilities**)

Set  $\hat{\rho}_i = \sum_{j \in [M]} [B]_{i,j}$ . Define  $\bar{\rho}_k = \frac{1}{M} e^k$ . Partition the vocabulary  $\mathcal{M}$  into:

$$\hat{\mathcal{I}}_0 = \{i : \hat{\rho}_i < \bar{\rho}_1\}, \text{ and } \hat{\mathcal{I}}_k = \{i : \bar{\rho}_k \leq \hat{\rho}_i \leq \bar{\rho}_{k+1}\}, \text{ for } k = 1 : \log M.$$

Sort the  $M$  words according to  $\hat{\rho}_i$  in ascending order.

Set  $\widehat{W}_k = \sum_{i \in \hat{\mathcal{I}}_k} \hat{\rho}_i$  and  $\widehat{M}_k = |\hat{\mathcal{I}}_k|$ . Set the block diagonal matrix

$$D_S = \begin{bmatrix} \bar{\rho}_1^{-1/2} I_{\widehat{M}_1} & & \\ & \ddots & \\ & & \bar{\rho}_{\log M}^{-1/2} I_{\widehat{M}_{\log M}} \end{bmatrix}. \quad (7)$$

Step 2. (**Estimate dictionary span in each bin**)

For each bin  $\hat{\mathcal{I}}_k$ , if  $\widehat{W}_k < \epsilon_0 e^{-k}$ , set  $\hat{V}_k = 0$ ; else consider diagonal block  $B_k = [B]_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k}$ :

- (a) (**Regularization**): Set  $d_k^{\max} = \widehat{W}_k \bar{\rho}_k$ . If a row/column of  $B_k$  has sum larger than  $2d_k^{\max}$ , set the entire row/column to 0. Denote the regularized block by  $\tilde{B}_k$ .
- (b) (**R-SVD**): Let the columns of  $\hat{V}_k$  denote the leading  $R$  singular vectors of  $\tilde{B}_k$ .

Step 3. (**Estimate dictionary span and an  $\ell_1$  estimator  $\hat{B}_2$** ) Set the projection matrix

$$\text{Proj}_{\hat{V}} = \begin{bmatrix} \text{Proj}_{\hat{V}_1} & & \\ & \ddots & \\ & & \text{Proj}_{\hat{V}_{\log M}} \end{bmatrix}. \quad (8)$$

- (a) (**Regularization**): For each word  $i$  in bin  $\hat{\mathcal{I}}_k$ , if the corresponding row in  $B$  has sum larger than  $2\bar{\rho}_k$ , set the entire row and column to zero. Denote the regularized average bigram matrix by  $\tilde{B}$ .
- (b) (**R-SVD**): Set  $\hat{B}_0$  to be the rank- $R$  truncated SVD of matrix  $\text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}}$ .  
Let the columns of  $\hat{V}$  denote the leading  $R$  singular vectors of  $\hat{B}_0$ .

Step 4. (**Refinement to get  $\ell_1$  estimator**)

Repeat the regularization in Step 3 on  $B$ , let  $\tilde{B}$  denote regularized average bigram matrix.

Set  $Y = (\hat{V}^\top D_S \tilde{B} D_S \hat{V})^{-1/2} (\hat{V}^\top D_S \tilde{B} D_S)$ , Set  $\hat{B} = D_S^{-1} Y Y^\top D_S^{-1}$ .

**Return**  $\hat{B}$  and  $\hat{V}$ .

**Algorithm 2:** Rank  $R$  algorithm

## References

- [1] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.
- [2] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.
- [3] Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2016.
- [4] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. In *Conference on Learning Theory (COLT)*, 2011.
- [5] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive classification and closeness testing. In *Conference on Learning Theory (COLT)*, 2012.
- [6] Anima Anandkumar, Yi kai Liu, Daniel J. Hsu, Dean P Foster, and Sham M Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25*. 2012.
- [7] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.
- [8] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization—provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012.
- [9] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models—going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012.
- [10] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *arXiv preprint arXiv:1502.03520*, 2015.
- [11] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR*, abs/1502.03520, 2015.
- [12] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1), 2013.
- [13] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Symposium on Theory of Computing (STOC)*, pages 381–390, 2004.
- [14] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.
- [15] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 594–603. ACM, 2014.

- [16] B. Bhattacharya and G. Valiant. Testing closeness with unequal sized samples. In *Neural Information Processing Systems (NIPS) (to appear)*, 2015.
- [17] L. Birge. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987.
- [18] J. T. Chang. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996.
- [19] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery. *arXiv preprint arXiv:1501.05021*, 2015.
- [20] Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.
- [21] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.
- [22] Joel Friedman, Jeff Kahn, and Endre Szemerédi. On the second eigenvalue of random regular graphs. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 587–598. ACM, 1989.
- [23] Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the Symposium on Theory of Computing, STOC 2015,*, 2015.
- [24] Peter W Glynn. Upper bounds on poisson tail probabilities. *Operations research letters*, 6(1):9–14, 1987.
- [25] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. In *Technical Report TR00-020, Electronic Colloquium on Computational Complexity*, 2000.
- [26] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.
- [27] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.
- [28] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.
- [29] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.
- [30] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.
- [31] Raghunandan H Keshavan, Sewoong Oh, and Andrea Montanari. Matrix completion from a few entries. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 324–328. IEEE, 2009.

- [32] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.
- [33] Can M Le, Elizaveta Levina, and Roman Vershynin. Sparse random graphs: regularization and concentration of the laplacian. *arXiv preprint arXiv:1502.03049*, 2015.
- [34] Can M Le and Roman Vershynin. Concentration and regularization of random graphs. *arXiv preprint arXiv:1506.00669*, 2015.
- [35] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*. 2014.
- [36] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.
- [37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [38] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.
- [39] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. *Annals of Applied Probability*, 16(2):583–614, 2006.
- [40] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.
- [41] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014.
- [42] S. on Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1193–1203, 2014.
- [43] L. Paninski. Estimating entropy on  $m$  bins given fewer than  $m$  samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004.
- [44] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.
- [45] Karl Stratos, Michael Collins, and Daniel Hsu. Model-based word embeddings from decompositions of count matrices. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 2015.
- [46] Karl Stratos, Michael Collins Do-Kyum Kim, and Daniel Hsu. A spectral algorithm for learning class-based n-gram models of natural language. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.



- [47] G. Valiant and P. Valiant. Estimating the unseen: an  $n/\log n$ -sample estimator for entropy and support size, shown optimal via new clts. In *Symposium on Theory of Computing (STOC)*, 2011.
- [48] G. Valiant and P. Valiant. The power of linear estimators. In *Symposium on Foundations of Computer Science (FOCS)*, 2011.
- [49] G. Valiant and P. Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Neural Information Processing Systems (NIPS)*, 2013.
- [50] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 51–60, 2014.
- [51] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.
- [52] Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block model.

### 3 Rank 2 Algorithm

Given  $N$  samples, the goal is to estimate the word marginal vector  $\rho$  as well as the dictionary separation vector  $\Delta$  up to constant accuracy in  $\ell_1$  norm. We denote the estimates by  $\hat{\rho}$  and  $\hat{\Delta}$ . Also, we estimate the underlying probability matrix  $\mathbb{B}$  with  $\hat{B} = \hat{\rho}\hat{\rho}^\top + \hat{\Delta}\hat{\Delta}^\top$ . Note that since  $\|\Delta\|_1 \leq \|\rho\|_1 = 1$ , constant  $\ell_1$  norm accuracy in  $\hat{\rho}$  and  $\hat{\Delta}$  immediately lead to constant accuracy of  $\hat{B}$  also in  $\ell_1$  norm.

In this section, we prove Theorem 3.1 and Theorem 3.2 about the correctness of the 2 Phases of Algorithm 1, the detailed proofs are provided in Section A in the appendix.

Throughout the section, we denote the ratio between sample size and the vocabulary size by

$$d_0 = N/M, \quad (9)$$

and we assume that  $d_0$  is lower bounded by a large constant such that

$$d_0 / \log d_0 > \epsilon_0^{-4}.$$

**Theorem 3.1** (Linear sample complexity of Rank 2 algorithm). *Fix  $\epsilon_0$  to be a small constant. Given  $N = \Theta(M)$  samples, with large probability, Phase I of Algorithm 1 estimates  $\rho$  and  $\Delta$  with accuracy:*

$$\|\hat{\rho} - \rho\|_1 < \epsilon_0, \quad \|\hat{\Delta} - \Delta\|_1 < \epsilon_0, \quad \|\hat{B} - \mathbb{B}\|_1 < \epsilon_0.$$

Under the separation assumptions of  $\Delta$ , we can refine the estimation to achieve arbitrary  $\epsilon$  accuracy.

**Theorem 3.2** (Refinement of Rank 2 algorithm). *Assume that  $\mathbb{B}$  satisfies the  $\Omega(1)$  separation assumption. Given  $N$  samples, with probability at least  $(1 - \delta)$ , Phase II of our Algorithm 1 estimates  $\rho$  and  $\Delta$  up to accuracy in  $\ell_1$  norm:*

$$\|\hat{\rho} - \rho\|_1 < \sqrt{M/\delta N}, \quad \|\hat{\Delta} - \Delta\|_1 < \sqrt{M/\delta N}, \quad \|\hat{B} - \mathbb{B}\|_1 = O(\sqrt{M/\delta N}).$$

First, we show that it is easy to estimate the marginal probability vector  $\rho$  up to constant accuracy.

**Lemma 3.3** (Estimate the word marginal probability  $\rho$ ). *Given the average empirical count matrix  $B$ , we estimate the marginal probabilities by:*

$$\hat{\rho}_i = \sum_{j \in \mathcal{M}} B_{i,j}. \quad (10)$$

*With probability at least  $(1 - \delta)$ , we can bound the estimation accuracy by:*

$$\|\hat{\rho} - \rho\|_1 \leq \frac{1}{\sqrt{d_0 \delta}}. \quad (11)$$

The hard part is to estimate the separation vector  $\Delta$  with linear number of sample counts, namely when  $d_0 = \Theta(1)$ . Recall that in the linear sample size regime, naively taking the rank-1 truncated SVD of  $(B - \hat{\rho}\hat{\rho}^\top)$  fails to reveal any information about  $\Delta\Delta^\top$ , since the leading eigenvectors of  $B$  are dominated by the statistical noise of the sampling words with large marginal. Algorithm 1 achieves this with delicate steps. The organization of this section is as follows:

1. Section 3.1 introduces the binning argument and the necessary notations for the rest of the section. We group the  $M$  words into bins according to the empirical marginal probabilities, i.e.  $\hat{\rho}_i$ 's. We call a bin “heavy” or “light” according to the marginal probability of a typical word in that bin.

2. Section 3.2 analyzes how to estimate the entries of  $\Delta$  restricted to different empirical bins (up to some common sign flip). To achieve this, for the heaviest bin where words' marginals are in the order of  $\Omega(\log M/M)$ , we can simply apply truncated SVD to the properly scaled diagonal block of the empirical average matrix  $B$ . For all other empirical bins, we examine the corresponding diagonal blocks in  $B$ . The main challenge here is to deal with the spillover effect due to inexact binning, and Lemma 4.1 shows that with high probability, such spillover effect is very small *for all bins* with high probability. Then we leverage the clever proof techniques from [34] to show that given small spillover effect, we can first regularize each diagonal block and then apply truncated SVD to estimate the segments of separation vector.
3. Section 3.3 shows how to stitch the segments of estimates for  $\Delta$  across different bins.
4. Section 3.4 shows that built upon the initialization, if the dictionary further satisfies certain separation condition, we can refine the estimation to improve its dependence on target accuracy  $\epsilon$  to meet the information theoretic lower bound.

### 3.1 Binning

In order to estimate the separation vector  $\Delta$ , instead of tackling the empirical count matrix  $B$  as a whole, we focus on its diagonal blocks and analyze the spectral concentration restricted to each block separately, using the fact that the entries  $\mathbb{B}_{i,j}$  restricted to each diagonal block are roughly uniform.

For any set of words  $\mathcal{I}$ , we use  $B_{\mathcal{I},\mathcal{I}}$  to denote the diagonal block of  $B$  whose row and column indices are in the set  $\mathcal{I}$ . When restricting to the diagonal block, the rank 2 decomposition of the expected matrix is given by  $\mathbb{B}_{\mathcal{I},\mathcal{I}} = \rho_{\mathcal{I}}\rho_{\mathcal{I}}^\top + \Delta_{\mathcal{I}}\Delta_{\mathcal{I}}^\top$ .

**Empirical binning** We partition the vocabulary  $\mathcal{M}$  according to the empirical marginal  $\hat{\rho}$  in (10):

$$\hat{\mathcal{I}}_0 = \left\{ i : \frac{\epsilon_0}{M} \leq \hat{\rho}_i < \frac{1}{M} \right\}, \quad \hat{\mathcal{I}}_k = \left\{ i : \frac{e^{k-1}}{M} \leq \hat{\rho}_i < \frac{e^k}{M} \right\}, \quad \hat{\mathcal{I}}_{\log} = \left\{ i : \frac{\log M}{M} \leq \hat{\rho}_i \right\}. \quad (12)$$

We call this *empirical binning* to emphasize the dependence on the empirical estimator  $\hat{\rho}$ , which is a random variable built from the first batch of  $N$  sample counts. We call  $\hat{\mathcal{I}}_0$  the *lightest empirical bin*, and  $\hat{\mathcal{I}}_{\log}$  the *heaviest empirical bin*, and  $\hat{\mathcal{I}}_k$  for  $1 \leq k \leq \log \log M$  the *moderate empirical bins*.

For the analysis, we further define the *exact bins* according to the exact marginal probabilities:

$$\mathcal{I}_0 = \left\{ i : \frac{\epsilon_0}{M} \leq \rho_i < \frac{1}{M} \right\}, \quad \mathcal{I}_k = \left\{ i : \frac{e^{k-1}}{M} \leq \rho_i < \frac{e^k}{M} \right\}, \quad \mathcal{I}_{\log} = \left\{ i : \frac{\log M}{M} \leq \rho_i \right\}. \quad (13)$$

Note that since the target accuracy of Phase I is a small constant  $\epsilon_0$ , we can safely discard all the words with marginals less than  $\epsilon_0/M$  as that incurs an  $\ell_1$  error only in the order of  $O(\epsilon_0)$ .

**Spillover effect** As  $N$  increases asymptotically, we will have  $\hat{\mathcal{I}}_k$  coincides with  $\mathcal{I}_k$  for every bin. However, in the linear regime where  $N = \Theta(M)$ , binning is inexact and we have the following two *spillover effects*:

1. Words from a heavy bin  $\mathcal{I}_{k'}$ , for  $k'$  much larger than  $k$ , are placed in a empirical bin  $\hat{\mathcal{I}}_k$ ;
2. Words from bin  $\mathcal{I}_k$  escape from the corresponding empirical bin  $\hat{\mathcal{I}}_k$ .

The hope, that we can have good spectral concentration in each diagonal block  $B_{\hat{\mathcal{I}}_k, \hat{\mathcal{I}}_k}$ , crucially relies on the fact that the entries  $\mathbb{B}_{i,j}$  restricted to this block are roughly uniform. However, the hope may be ruined by the spillover effects. Next, we show that with high probability the spillover effects are small for all bins with large probability mass:

1. In each empirical bin  $\widehat{\mathcal{I}}_k$ , the total probability mass of heavy words from  $\cup_{\{k':k'>k+1\}}\mathcal{I}_{k'}$  is small and in the order of  $O(e^{-e^k d_0/2})$  (see Lemma 4.1).
2. Most words of  $\mathcal{I}_k$  stays within the nearest empirical bins, namely  $\cup_{\{k':k-1\leq k'\leq k+1\}}\widehat{\mathcal{I}}_{k'}$ , (see Lemma 3.15).

**Notations** To analyze the spillover effects, we define some additional quantities.

We define the total marginal probability mass in the empirical bins to be:

$$W_k = \sum_{i \in \widehat{\mathcal{I}}_k} \rho_i, \quad (14)$$

and let  $M_k = |\widehat{\mathcal{I}}_k|$  denote the total number of words in the empirical bin. We also define  $\widehat{W}_k = \sum_{i \in \widehat{\mathcal{I}}_k} \widehat{\rho}_i$ .

We use  $\widehat{\mathcal{J}}_k$  to denote the set of spillover words into the empirical bin  $\widehat{\mathcal{I}}_k$ :

$$\widehat{\mathcal{J}}_k = \widehat{\mathcal{I}}_k \cap (\cup_{\{k':k'>k+1\}}\mathcal{I}_{k'}), \quad (15)$$

and let  $\widehat{\mathcal{L}}_k$  denote the “good words” in the empirical bin  $\widehat{\mathcal{I}}_k$ :

$$\widehat{\mathcal{L}}_k = \widehat{\mathcal{I}}_k \setminus \widehat{\mathcal{J}}_k. \quad (16)$$

We also denote the total marginal probability mass of the heavy spillover words  $\widehat{\mathcal{J}}_k$  by:

$$\overline{W}_k = \sum_{i \in \widehat{\mathcal{J}}_k} \rho_i. \quad (17)$$

Note that these quantities are random variables determined by the randomness of the first batch of  $N$  samples, in the binning step. We fix the binning when considering the empirical count matrix  $B$  (with independent batches of samples) in the other steps of the algorithm.

Define the upper bound of the “typical” word marginal in the  $k$ -th empirical bin to be:

$$\bar{\rho}_k = e^{\tau+1}/M,$$

Recall that we have  $\sum_k \mathbb{B}_{i,k} = w_p p + w_q q$  and we assume  $w_p, w_q \geq C_w = \Omega(1)$ . We can bound each entry in  $\mathbb{B}$  by the product of marginals probabilities as

$$\mathbb{B}_{i,j} \leq \frac{2}{C_w^2} \rho_i \rho_j, \quad \forall i, j.$$

Let  $d_k^{\max}$  denote the expected max row/column sum of the diagonal block  $\mathbb{B}_{\widehat{\mathcal{I}}_k, \widehat{\mathcal{I}}_k}$ :

$$d_k^{\max} =: M_k \max_{i,j \in \mathcal{I}_k} \mathbb{B}_{i,j} = 2M_k \bar{\rho}_k^2 / C_w^2. \quad (18)$$

### 3.2 Estimate segments of $\Delta$

**Heaviest empirical bin** First, we show that the empirical marginal probabilities of words in the heaviest bin concentrate much better than what Lemma 3.3 implies.

**Lemma 3.4** (Concentration of marginal probabilities in the heaviest bin). *With high probability, for all the words with marginal probability  $\rho_i \geq \epsilon_0 \log(M)/M$ , for some universal constant  $C_1, C_2$ ,*

$$C_1 \leq \widehat{\rho}_i / \rho_i \leq C_2. \quad (19)$$

Lemma 3.4 says that we can estimate the marginal probabilities for every words in the heaviest bin with constant multiplicative accuracy. It also suggests that we do not need to worry about the words from  $\mathcal{I}_{\log}$  get spilled over into much lighter bins.

The next lemmas shows that with proper scaling, we can apply truncated SVD to the diagonal block to estimate the entries of separation vector  $\Delta$  restricted to the empirical heaviest bin.

**Lemma 3.5** (Estimate  $\Delta$  restricted to the heaviest empirical bin). *Suppose that  $\widehat{W}_{\log} = \sum \widehat{\rho}_{\widehat{\mathcal{I}}_{\log}} > \epsilon_0$ . Define  $\widehat{D}_{\widehat{\mathcal{I}}_{\log}} = \text{Diag}(\widehat{\rho}_{\widehat{\mathcal{I}}_{\log}})$ . Consider  $B_{\widehat{\mathcal{I}}_{\log}, \widehat{\mathcal{I}}_{\log}}$ , the diagonal block corresponding to  $\widehat{\mathcal{I}}_{\log}$ . Let  $E$  be the rank 1 truncated SVD of  $\widehat{D}_{\widehat{\mathcal{I}}_{\log}}^{-1/2} (B_{\widehat{\mathcal{I}}_{\log}, \widehat{\mathcal{I}}_{\log}} - \widehat{\rho}_{\widehat{\mathcal{I}}_{\log}} \widehat{\rho}_{\widehat{\mathcal{I}}_{\log}}^\top) \widehat{D}_{\widehat{\mathcal{I}}_{\log}}^{-1/2}$ . Set  $v_{\log} = \widehat{D}_{\widehat{\mathcal{I}}_{\log}}^{1/2} E^{1/2}$ . With large probability, we can estimate the dictionary separation vector restricted to the heaviest empirical bin up to sign flip with accuracy:*

$$\min\{\|\Delta_{\widehat{\mathcal{I}}_{\log}} - v_{\log}\|_1, \|\Delta_{\widehat{\mathcal{I}}_{\log}} + v_{\log}\|_1\} = O\left(\min\left\{\frac{1/d_0^{1/2}}{\|\Delta_{\widehat{\mathcal{I}}_{\log}}\|_1}, 1/d_0^{1/4}\right\}\right). \quad (20)$$

The two cases in the above bound correspond to whether the separation is large or small, compared to the statistical noise from sampling, which is in the order  $1/d_0^{1/4}$ . If the bin contains a large separation, then the bound follows the standard Wedin's perturbation bound; if the separation is small, i.e.  $\|\Delta_{\widehat{\mathcal{I}}_{\log}}\|_1 \ll 1/d_0^{1/4}$ , then the bound  $1/d_0^{1/4}$  just corresponds to the magnitude of the statistical noise.

**Moderate empirical bins** In Lemma 4.1, we upper bound the spillover probability  $\overline{W}_k$  to show that the spillover effects are small for all the moderate bins. Given that, Lemma 3.7 and Lemma 3.8 show that we can first regularize each diagonal block and then apply truncated SVD to estimate the entries of the separation vector  $\Delta$  restricted to each bin.

**Lemma 3.6** (Bound spillover probabilities). *With high probability, for all empirical bins, we can bound  $\overline{W}_k$  defined in (17), the spillover probability from much heavier bins, by:*

$$\overline{W}_k \leq 2e^{-e^k d_0/2}. \quad (21)$$

Now consider  $B_{\widehat{\mathcal{I}}_k, \widehat{\mathcal{I}}_k}$ , the diagonal block corresponding to bin  $\widehat{\mathcal{I}}_k$ . We restrict attention to its spectral concentration on indices of  $\widehat{\mathcal{L}}_k$ , the set of “good words” defined in (16). To ensure the spectral concentration, we “regularize” it by removing the rows and columns with abnormally large sum. Recall that the expected row sum of the diagonal block without spillover is bounded by  $d_k^{\max}$  defined in (18). Let  $\widehat{\mathcal{R}}_k$  denote the indices of the rows and columns in  $B_{\widehat{\mathcal{I}}_k, \widehat{\mathcal{I}}_k}$  whose row sum or column sum are larger than  $2d_k^{\max}$ , namely

$$\widehat{\mathcal{R}}_k = \left\{i \in \widehat{\mathcal{I}}_k : \sum_{j \in \widehat{\mathcal{I}}_k} B_{i,j} > 2d_k^{\max} \text{ or } \sum_{j \in \widehat{\mathcal{I}}_k} B_{j,i} > 2d_k^{\max}\right\}. \quad (22)$$

Starting with  $\widetilde{B}_k = B_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}$ , we set all the rows and columns of  $\widetilde{B}_k$  indexed by  $\widehat{\mathcal{R}}_k$  to 0.

To make the operation of “regularization” more precise, we introduce some additional notations. Define  $\widetilde{\rho}_k$  to be a vector with the same length as  $\rho_{\widehat{\mathcal{I}}_k}$ , with the entries spillover words  $\widehat{\mathcal{J}}_k$  set to 0,

$$(\widetilde{\rho}_k)_i = \rho_i \mathbf{1}_{i \in \widehat{\mathcal{L}}_k}. \quad (23)$$

Similarly define vector  $\widetilde{\Delta}_k$  to be the separation vector restricted to the good words:

$$(\widetilde{\Delta}_k)_i = \Delta_i \mathbf{1}_{i \in \widehat{\mathcal{L}}_k}. \quad (24)$$

We define the matrix  $\tilde{\mathbb{B}}_k$  (of the same size as  $B_{\hat{\mathcal{I}}_k, \hat{\mathcal{I}}_k}$ ):

$$\tilde{\mathbb{B}}_k = \tilde{\rho}_k \tilde{\rho}_k^\top + \tilde{\Delta}_k \tilde{\Delta}_k^\top. \quad (25)$$

Note that by definition the rows and columns in  $\tilde{B}_k$  and  $\tilde{\mathbb{B}}_k$  that are zero-ed out do not necessarily coincide. However, the next lemma shows that  $\tilde{B}_k$  concentrates to  $\tilde{\mathbb{B}}_k$  in the spectral distance.

**Lemma 3.7** (Spectral concentration of diagonal blocks.). *Suppose that the marginal of the bin  $\hat{\mathcal{I}}_k$  is large enough  $W_k = \sum \rho_{\hat{\mathcal{I}}_k} > \epsilon_0 e^{-k}$ . With probability at least  $(1 - M_k^{-r})$ , for some universal constant  $r$ , we have*

$$\|\tilde{B}_k - \tilde{\mathbb{B}}_k\|_2 \leq C r^{1.5} \frac{\sqrt{N d_k^{\max} \log(N d_k^{\max})}}{N}. \quad (26)$$

*Proof.* Here we highlight the key steps of the proof, and defer the detailed proof to Section A.

In Figure 2, the rows and the columns of  $B_{\hat{\mathcal{I}}_k, \hat{\mathcal{I}}_k}$  are sorted according to the exact marginal probabilities of the words in ascending order, with the rows and columns set to 0 by regularization shaded. Consider the block decomposition according to the good words  $\hat{\mathcal{L}}_k$  and the spillover words  $\hat{\mathcal{J}}_k$ . We bound the spectral distance of the 4 blocks  $(A_1, A_2, A_3, A_4)$  separately. The bound for the entire matrix  $\tilde{B}_k$  is then an immediate result of triangle inequality.

For block  $A_1$  whose rows and columns all correspond to the “good words” with roughly uniform marginals, we show its concentration by applying the result in [34]. For block  $A_2$  and  $A_3$ , we show that after regularization the spectral norm of these two blocks are small. Intuitively, the expected row sums of block  $A_2$  are bounded by  $2d_k^{\max}$  and the expected column sums are bounded by  $2d_k^{\max} \frac{\bar{W}_k}{W_k} = O(1/N)$ , as a result of the bound on  $\bar{W}_k$  in Lemma 4.1. Thus the spectral norm of the block  $A_2$  is likely to be bounded by  $O(\sqrt{d_k^{\max}/N})$ . We show this rigorously with high probability arguments. Lastly for block  $A_4$ , which rows and columns all correspond to the spillover words. We show that the spectral norm of this block is very small, as a result of the small spillover marginal  $\bar{W}_k$ .  $\square$

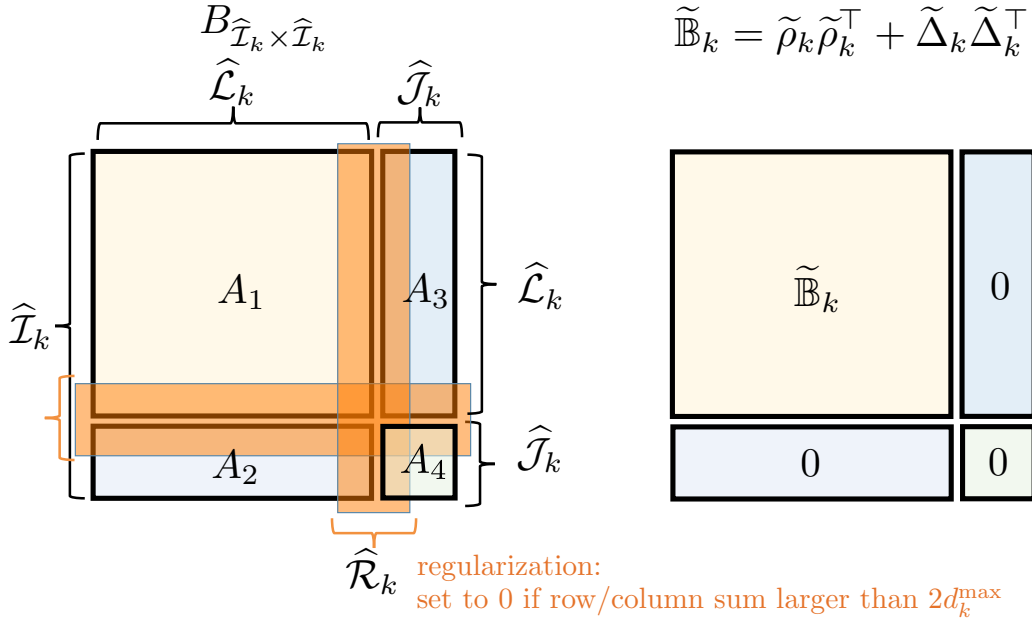


Figure 2: block decomposition of the diagonal block of  $B_{\hat{\mathcal{I}}_k, \hat{\mathcal{I}}_k}$  corresponding to  $\hat{\mathcal{I}}_k$ .

**Lemma 3.8** (Estimate the separation vector restricted to bins). *Suppose that  $W_k = \sum_{i \in \hat{\mathcal{I}}_k} \rho_i > C_1 e^{-k}$  for some fixed constant  $C_1 = \Omega(1)$ . Let  $v_k v_k^\top$  be the rank-1 truncated SVD of the matrix  $(\hat{B}_k - \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top)$ . With high probability, we have*

$$\min\{\|\tilde{\Delta}_k - v_k\|_2, \|\tilde{\Delta}_k + v_k\|_2\} = O\left(\min\left\{\frac{\sqrt{Nd_k^{\max} \log(Nd_k^{\max})}}{N} \frac{1}{\|\Delta_{\hat{\mathcal{I}}_k}\|_2}, \left(\frac{\sqrt{Nd_k^{\max} \log(Nd_k^{\max})}}{N}\right)^{1/2}\right\}\right). \quad (27)$$

**Claim 3.9** (Estimate the separation vector restricted to the lightest bin). *Setting  $\hat{\Delta}_{\hat{\mathcal{I}}_0} = 0$  only incurs a small constant error:*

$$\|\Delta_{\hat{\mathcal{I}}_0}\|_1 \leq \|\rho_{\hat{\mathcal{I}}_0}\|_1 \leq \|\rho_{\mathcal{I}_0}\|_1 + \bar{W}_0 \leq \frac{\epsilon_0}{M} M + 2e^{-d_0/2} = O(\epsilon_0),$$

where we used the assumption that  $d_0/\log d_0 \geq \epsilon_0^{-4}$ .

### 3.3 Stitch the segments of $\hat{\Delta}$

Given  $v_k$  for all  $k$  as estimation for  $\Delta_{\hat{\mathcal{I}}_k}$ 's up to sign flips. Fix  $k^*$  to be one good bin (with large bin marginal and large separation). Partition the words into two groups  $\mathcal{I}_{k^*}^+ = \{i : i \in \mathcal{I}_{k^*} : \hat{\Delta}_i > 0\}$  and  $\mathcal{I}_{k^*}^- = \mathcal{I}_{k^*} \setminus \mathcal{I}_{k^*}^+$ . Without loss of generality assume that  $\sum_{i \in \mathcal{I}_{k^*}^+} \hat{\Delta}_i \geq \sum_{i \in \mathcal{I}_{k^*}^-} \hat{\Delta}_i$ . We set  $\hat{\Delta}_{\hat{\mathcal{I}}_{k^*}} = v_{k^*}$ . For all other good bins  $k$ , we similarly define  $\mathcal{I}_k^+$  and  $\mathcal{I}_k^-$ . The next claim shows how to determine the relative sign flip of  $v_{k^*}$  and  $v_k$ .

**Claim 3.10** (Pairwise comparison of bins to fix sign flips). *For all good bins  $k \in \mathcal{G}$ , we can fix the sign flip to be  $\hat{\Delta}_{\hat{\mathcal{I}}_k} = v_k$  if:*

$$\frac{\sum_{i \in \mathcal{I}_{k^*}^+, j \in \mathcal{I}_k^+} [B]_{i,j}}{\sum_{i \in \mathcal{M}, j \in \mathcal{I}_k^+} [B]_{i,j}} > \frac{\sum_{i \in \mathcal{I}_{k^*}^+, j \in \mathcal{I}_k^-} [B]_{i,j}}{\sum_{i \in \mathcal{M}, j \in \mathcal{I}_k^-} [B]_{i,j}},$$

and  $\hat{\Delta}_{\hat{\mathcal{I}}_k} = -v_k$  otherwise.

*Proof.* This claim is straightforward. When restricted to the good bins, the estimates  $v_k$  are accurate enough. We can determine that the sign flips of  $k^*$  and  $k$  are consistent if and only if the conditional distribution of the two word tuple  $(x, y) \in \mathcal{M}^2$  satisfies  $\Pr(x \in \mathcal{I}_{k^*}^+ | x \in \mathcal{I}_k^+) > \Pr(x \in \mathcal{I}_{k^*}^+ | x \in \mathcal{I}_k^-)$ , and we should revert  $v_k$  otherwise.  $\square$

Concatenate the segments of  $\hat{\Delta}$ , we can bound the overall estimation error of the separation vector.

**Lemma 3.11** (Estimate separation vector in Phase I). *For a fixed small constant  $\epsilon_0 = O(1)$ , if  $d_0/\log(d_0) \geq \epsilon_0^{-4}$ , with large probability, Phase I of Algorithm 1 estimates the separation vector  $\Delta$  with constant accuracy in  $\ell_1$  norm:*

$$\|\hat{\Delta} - \Delta\| = O(\epsilon_0).$$

This concludes the proof for Theorem 3.1.

### 3.4 Refinement

**Construct an anchor partition** Imagine that we have a way to group the  $M$  words in the vocabulary into a new vocabulary with a *constant* number of superwords. The new probability matrix is obtained by summing over the rows and columns of the matrix  $\mathbb{B}$  according to the grouping. We similarly define marginal vector  $\rho_A$  and separation vector  $\Delta_A$  over the superwords. If we group the words in a way such that the dictionary over the superwords is still well separated, then with  $N = \Omega(M)$  samples we can estimate the constant dimensional  $\rho_A$  and  $\Delta_A$  to arbitrary accuracy. Such estimates provide us some crude and global information about the true original dictionary. Now sum the probability matrix only over the rows accordingly, the expectation can be factorized as  $\rho_A \rho^\top + \Delta_A \Delta^\top$ . Therefore, given accurate estimates of  $\rho_A$  and  $\Delta_A$ , obtaining refined estimation  $\hat{\rho}$  and  $\hat{\Delta}$  is as simple as solving a least square problem.

**Definition 3.12** (Anchor partition). *Consider a partition of the vocabulary  $[M]$  into  $(\mathcal{A}, \mathcal{A}^c)$ . denote  $\rho_{\mathcal{A}} = \sum_{i \in \mathcal{A}} \rho_i$  and  $\Delta_{\mathcal{A}} = \sum_{i \in \mathcal{A}} \Delta_i$ . We call it an anchor partition if for some constant  $C_A = \Omega(1)$ ,*

$$\text{cond} \left( \begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix} \right) \leq C_A. \quad (28)$$

If the dictionary is well separated  $\|\Delta\|_1 = \Omega(1)$ , it is feasible to find an anchor partition. Moreover, we will show that we can use the estimator  $\hat{\Delta}$  obtained in Phase I to construct such an anchor partition easily. The next lemma states a sufficient condition for constructing an anchor partition.

**Lemma 3.13** (Sufficient condition for constructing an anchor partition). *Let  $\Delta_{\mathcal{I}}$  be the vector of  $\Delta$  restricted to a set of words  $\mathcal{I}$ . Suppose that  $\|\Delta_{\mathcal{I}}\|_1 \geq C \|\Delta\|_1$  for some constant  $C = \Omega(1)$ , and that for some constant  $C' \leq \frac{1}{3}C$ , we can estimate  $\Delta_{\mathcal{I}}$  up to precision:*

$$\|\hat{\Delta}_{\mathcal{I}} - \Delta_{\mathcal{I}}\|_1 \leq C' \|\Delta_{\mathcal{I}}\|_1. \quad (29)$$

Denote  $\hat{\mathcal{A}} = \{i \in \mathcal{I} : \hat{\Delta}_i > 0\}$ . We have that  $(\hat{\mathcal{A}}, \mathcal{M} \setminus \hat{\mathcal{A}})$  forms an anchor partition defined in 3.12.

**Definition 3.14** (Good bins). *Denote the dictionary separation restricted to the “good words” in each empirical bin  $\hat{\mathcal{I}}_k$  by:*

$$S_k =: \sum_{i \in \hat{\mathcal{I}}_k} |\Delta_i| = \|\tilde{\Delta}_k\|_1. \quad (30)$$

Fix constants  $C_1 = C_2 = \frac{1}{24} \|\Delta\|_1 = \Omega(1)$ . We call bin  $\hat{\mathcal{I}}_k$  a “good bin” if it satisfies that:

1. the marginal probability of the bin  $W_k \geq C_1 e^{-k}$ .
2. the ratio between the separation and the marginal probability of the bin satisfies  $\frac{S_k}{2W_k} \geq C_2$ .

Let  $\mathcal{G}$  denote the set of all the good bins. Next lemma shows that a constant fraction of total probability mass is contained in good bins.

**Lemma 3.15** (Total mass in good bins). *With high probability, we can bound the total marginal probability mass in the “good bins” by:*

$$\sum_{k \in \mathcal{G}} W_k \geq \|\Delta\|_1 / 12. \quad (31)$$

This implies a bound of total separation contained in all the good words of the good bins:

$$\sum_{i \in \hat{\mathcal{I}}_k, k \in \mathcal{G}} |\Delta_i| = \sum_{k \in \mathcal{G}} S_k \geq 2C_2 \sum_{k \in \mathcal{G}} W_k \geq \frac{1}{24} (\|\Delta\|_1)^2 = \Omega(1). \quad (32)$$



**Lemma 3.16** (Estimate the separation vector restricted to good bins). *If the empirical bin  $\hat{\mathcal{I}}_k$  is a good bin, with high probability, the estimate  $\hat{\Delta}_{\hat{\mathcal{I}}_k}$  from Phase I (Lemma 3.8), for the separation vector restricted to the bin satisfies:*

$$\|\hat{\Delta}_{\hat{\mathcal{I}}_k} - \tilde{\Delta}_k\|_1 \leq \frac{1}{\sqrt{d_0}} \|\Delta_{\hat{\mathcal{I}}_k}\|_1. \quad (33)$$

The above two lemmas suggest that we can focus on the “good words” in the “good bins”, namely  $\mathcal{I} = \cup_{k \in \mathcal{G}} \hat{\mathcal{I}}_k$ . Lemma 3.15 showed the separation contained in  $\mathcal{I}$  is at least  $\sum_{k \in \mathcal{G}} S_k = C \|\Delta\|_1$  for some  $C = \Omega(1)$ ; Lemma 3.16 showed that with linear number of samples we can estimate  $\Delta$  restricted to  $\mathcal{I}$  up to constant accuracy. Therefore by Lemma 3.13 we can construct a valid anchor partition  $(\mathcal{A}, \mathcal{M} \setminus \mathcal{A})$  by setting:  $\mathcal{A} = \{i : \hat{\Delta}_i > 0, \text{ for } i \in \hat{\mathcal{I}}_k, k \in \mathcal{G}\}$ .

Ideally, we want to restrict to the “good words” and set the anchor partition to be  $\{i : \hat{\Delta}_i > 0, \text{ for } i \in \tilde{\mathcal{L}}_k, k \in \mathcal{G}\}$ , but we cannot distinguish the “good words” from spillover words. However, the bound on the total marginal of spillover  $\sum_k \bar{W}_k = O(e^{-d_0/2})$  guarantees that even if we mis-classify all the spillover words, the construction is still a valid anchor partition.

**Estimate the anchor matrix** Given the two superwords  $(\mathcal{A}, \mathcal{M} \setminus \mathcal{A})$  from the anchor partition, define the  $2 \times 2$  matrix  $D_{\mathcal{A}} = \begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix}$  to be the anchor matrix. To estimate the two scalars  $\rho_{\mathcal{A}}$  and  $\Delta_{\mathcal{A}}$ , we apply the standard concentration bound and argue that with high probability,

$$\left\| \begin{bmatrix} \sum_{i \in \mathcal{A}, j \in \mathcal{A}} [B_N]_{i,j} & \sum_{i \in \mathcal{A}, j \in \mathcal{A}^c} [B_N]_{i,j} \\ \sum_{i \in \mathcal{A}^c, j \in \mathcal{A}} [B_N]_{i,j} & \sum_{i \in \mathcal{A}^c, j \in \mathcal{A}^c} [B_N]_{i,j} \end{bmatrix} - D_{\mathcal{A}} D_{\mathcal{A}}^{\top} \right\| = O\left(\frac{1}{\sqrt{N}}\right)$$

Recall that by anchor partition, we have  $|\Delta_{\mathcal{A}}| = \Omega(1)$ . Thus we can estimate  $\rho_{\mathcal{A}}$  and  $\Delta_{\mathcal{A}}$  to accuracy  $\frac{1}{\sqrt{N}}$ . Since  $N = \Omega(M)$  is asymptotically large, we essentially obtain precisely the anchor matrix  $D_{\mathcal{A}}$ .

**Use anchor matrix to refine estimation** Now given an anchor partition of the vocabulary  $(\mathcal{A}, \mathcal{A}^c)$ , and given the exact anchor matrix  $D_{\mathcal{A}}$  which has  $\Omega(1)$  condition number, refining the estimation of  $\rho_i$  and  $\Delta_i$  for each  $i$  is very easy and achieves optimal rate.

**Lemma 3.17** (Refine estimation). *With probability at least  $1 - \delta$ , Phase II of Algorithm 1 outputs estimates  $\hat{\rho}$  and  $\hat{\Delta}$  such that*

$$\|\hat{\rho} - \rho\| < \sqrt{1/\delta N}, \quad \|\hat{\Delta} - \Delta\| < \sqrt{1/\delta N}.$$

The above lemma implies the  $\ell_1$  norm accuracy for Theorem 3.2:

$$\|\hat{\rho} - \rho\|_1 < \sqrt{M/\delta N}, \quad \|\hat{\Delta} - \Delta\|_1 < \sqrt{M/\delta N}.$$

## 4 Rank $R$ Algorithm

In this section, we examine each step of Algorithm 2 to prove Theorem 1.2 and Theorem 1.3. Recall that we are given 4 independent batches of  $N$  samples, with which we construct 4 independent empirical bigram matrix  $B = \mathbb{B} + E$  where the noise matrix  $E$  are independent and identical copies of sampling noise  $E$ . In each of the 4 steps of the algorithm, an independent and identical copy of the bigram matrix is used. We omit the index  $i = 1, \dots, 4$  for notation brevity.

## 4.1 Binning

We focus on the symmetric case where the rank  $R$  probability matrix is parameterized as  $\mathbb{B} = PWP^\top$ , and  $W$  is a PSD matrix. The algorithm and analysis can be easily extended to deal with more general case. We define the weight  $w_r = \sum_i W_{r,i}$ . We assume that the weights are lower bounded by

$$w_{\min} = \min_r w_r \geq C_w = \Omega(1).$$

The marginal probability is given by

$$\rho_i = \sum_k \mathbb{B}_{i,k} = \sum_r w_r p_i^{(r)}.$$

Note that each entry of the probability matrix  $\mathbb{B}$  can be bounded by the product of the corresponding marginal probability as below:

$$\mathbb{B}_{i,j} = \sum_{s,t} W_{s,t} p_i^{(s)} p_j^{(t)} \leq \sum_{s,t} W_{s,t} p_i^{(s)} \sum_{t'} p_j^{(t')} = \rho_i \sum_{t'} p_j^{(t')} \leq \frac{1}{w_{\min}} \rho_i \rho_j. \quad (34)$$

Again, binning according to the empirical marginal is given by:

$$\widehat{\mathcal{I}}_0 = \left\{ i : \frac{\epsilon_0}{M} \leq \widehat{\rho}_i < \frac{1}{M} \right\}, \quad \widehat{\mathcal{I}}_k = \left\{ i : \frac{e^{k-1}}{M} \leq \widehat{\rho}_i < \frac{e^k}{M} \right\}, \text{ for } k = 1 : \log M.$$

Let  $M_k = |\widehat{\mathcal{I}}_k|$  denote the number of words in bin  $\widehat{\mathcal{I}}_k$ .

The grouping of words according to the exact marginal probabilities is defined as:

$$\mathcal{I}_0 = \left\{ i : \frac{\epsilon_0}{M} \leq \rho_i < \frac{1}{M} \right\}, \quad \mathcal{I}_k = \left\{ i : \frac{e^{k-1}}{M} \leq \rho_i < \frac{e^k}{M} \right\}, \text{ for } k = 1 : \log M.$$

Define  $\bar{\rho}_k$  to be the typical marginal of a word in bin  $\mathcal{I}_k$ :

$$\bar{\rho}_k = e^k / M.$$

For  $i, j \in \mathcal{I}_k$ , we have  $\mathbb{B}_{i,j} \leq \bar{\rho}_k^2 / w_{\min}$ .

Due to the statistical noise of sampling,  $\widehat{\mathcal{I}}_k$  may contain words whose exact marginal is much larger than  $\bar{\rho}_k$ . The next lemma argues that such spillover effect is small.

**Lemma 4.1** (Spillover from heavier bins is small). *With high probability, for all empirical bins  $\widehat{\mathcal{I}}_k$ , we can bound the spillover probability from much heavier bins by:*

$$\overline{W}_k := \sum_{i \in \mathcal{I}_{k'} : k' > k + \tau} \rho_i \leq 2e^{-e^{\tau+k} d_0 / 2}. \quad (35)$$

**Definition 4.2** (Big bin). *An empirical bin  $\widehat{\mathcal{I}}_k$  is a big bin if*

$$W_k = \sum_{j \in \widehat{\mathcal{I}}_k} \rho_j > e^{-k}. \quad (36)$$

We know that a constant fraction of all the probability mass lies in such big bins. Moreover for  $d_0 \gg 1$ , we have  $W_k > e^{-k} \gg 2e^{-(k+\tau)d_0/2} \geq \overline{W}_k$ .

**Lemma 4.3** (Escaped probability mass). *With high probability, for all big bins, the mass that escapes from the bin is bounded by*

$$W_k^s =: \sum_{i \in \mathcal{I}_k, i \notin \widehat{\mathcal{I}}_k \text{ for } |k-k'| < \tau} \rho_i \leq 4W_k e^{-e^{k+\tau} d_0 / 2}.$$

## 4.2 Spectral concentration in diagonal blocks

Define the regularized probability matrix  $\tilde{\mathbb{B}}$  by setting the rows/columns corresponding to spillover words from much heavier bins to 0:

$$\tilde{\mathbb{B}} = \text{Diag}(\mathbf{1}[\rho_i < 2\bar{\rho}_k])\mathbb{B}\text{Diag}(\mathbf{1}[\rho_i < 2\bar{\rho}_k]) \quad (37)$$

Under the assumption that  $W$  is a PSD matrix, we define the  $M \times R$  matrix  $\mathbb{B}^{sqr t}$  and  $\tilde{\mathbb{B}}^{sqr t}$  to be:

$$\mathbb{B}^{sqr t} = PW^{1/2}, \quad \text{and} \quad \tilde{\mathbb{B}}^{sqr t} = \text{Diag}(\mathbf{1}[\rho_i < 2\bar{\rho}_k])PW^{1/2}. \quad (38)$$

Consider the diagonal block corresponding to the  $k$ -th empirical bin

$$B_k = [B]_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k}.$$

Similarly, we define the  $M_k \times R$  matrix restricting bin  $\hat{\mathcal{I}}_k$  as:

$$\mathbb{B}_{\hat{\mathcal{I}}_k}^{sqr t} = P_{\hat{\mathcal{I}}_k} W^{1/2}, \quad \text{and} \quad \tilde{\mathbb{B}}_{\hat{\mathcal{I}}_k}^{sqr t} = \text{Diag}(\mathbf{1}[\rho_i < 2\bar{\rho}_k])P_{\hat{\mathcal{I}}_k} W^{1/2},$$

We argue that, given  $N = \Omega(MR^2)$ , for each diagonal block  $B_k$ , Step 2 of Algorithm 2 finds a subspace  $\hat{V}_k$  correlated with  $\tilde{\mathbb{B}}_k^{sqr t}$ . We can bound  $\|\text{Proj}_{\hat{V}_k} \tilde{\mathbb{B}}_{\hat{\mathcal{I}}_k}^{sqr t} - \tilde{\mathbb{B}}_{\hat{\mathcal{I}}_k}^{sqr t}\|$  up to constant accuracy.

**Definition 4.4** (Expected row sum in diagonal blocks). *Recall that for  $i, j \in \mathcal{I}_k$ , we have  $\mathbb{B}_{i,j} \leq \bar{\rho}_k^2/w_{\min}$ . Define the maximal expected row sum of the diagonal block  $B_k$  to be:*

$$d_k^{max} = M_k \bar{\rho}_k^2 / w_{\min}. \quad (39)$$

Note that in the particular parameterization for the problem community detection with uniform marginal, we can simply define  $d_k^{max}$  to be  $1/M$  and thus get rid of the  $w_{\min}^{-1}$  dependence, and the rest of the analysis follows to recover the sample complexity result of  $N = \Theta(MR^2)$  in the community detection problem with  $R$  communities.

**Lemma 4.5** (Spectral concentration in each diagonal block). *Regularize the  $k$ -th diagonal block  $B_k$  by removing the rows/columns with sum larger than  $2d_k^{max}$ . Run rank  $R$  truncated SVD on the regularized block  $\tilde{B}_k$ . Let the columns of the  $M_k \times R$  matrix  $\hat{V}_k$  be the leading  $R$  singular vectors. Define  $\text{Proj}_{\hat{V}_k} = \hat{V}_k \hat{V}_k^\top$ . We have*

$$\|\text{Proj}_{\hat{V}_k} \tilde{\mathbb{B}}_{\hat{\mathcal{I}}_k}^{sqr t} - \tilde{\mathbb{B}}_{\hat{\mathcal{I}}_k}^{sqr t}\| = O\left(\frac{\sqrt{Nd_k^{max} \log Nd_k^{max}}}{N}\right)^{1/2}. \quad (40)$$

## 4.3 Low rank projection

In Step 3 of Algorithm 2, we “stitch” the subspaces  $\hat{V}_k$  for each bin  $\hat{\mathcal{I}}_k$  learned in Step 2 to get an estimate for the column span of the entire matrix.

Define the diagonal matrix  $D_S$  of dimension  $M \times M$  to be:

$$D_S = \begin{bmatrix} \bar{\rho}_1^{-1/2} I_{\hat{M}_1} & & \\ & \ddots & \\ & & \bar{\rho}_{\log M}^{-1/2} I_{\hat{M}_{\log M}} \end{bmatrix}. \quad (41)$$

Define  $\text{Proj}_{\widehat{V}}$  to be the block diagonal projection matrix which projects an  $M \times M$  matrix to a subspace  $\widehat{V}$  of dimension at most  $R \log M$ :

$$\text{Proj}_{\widehat{V}} = \begin{bmatrix} \text{Proj}_{\widehat{V}_1} & & \\ & \ddots & \\ & & \text{Proj}_{\widehat{V}_{\log M}} \end{bmatrix}. \quad (42)$$

Now consider the empirical average bigram  $B$  with the 3rd batch of samples. We regularize the entire matrix in the following way. For each row in  $B$ , if the word  $i$  is in bin  $\widehat{\mathcal{T}}_k$  as defined in Step 1, and if the corresponding row has sum larger than  $2\bar{\rho}_k$ , we set the entire row and column to zero. Let  $\widetilde{B}$  denote the regularized matrix.

**Lemma 4.6.** *Let  $\widehat{B}_1$  denote the rank  $R$  truncated SVD of  $\text{Proj}_{\widehat{V}} D_S \widetilde{B} D_S \text{Proj}_{\widehat{V}}$ . With large probability, we can bound the spectral distance between  $\widehat{B}_1$  and  $D_S \widetilde{B} D_S$  by:*

$$\|\widehat{B}_1 - D_S \widetilde{B} D_S\| = O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/M}\right)^{1/4}. \quad (43)$$

**Lemma 4.7.** *Let  $\widehat{B}_2 = D_S^{-1} \widehat{B}_1 D_S^{-1}$ , we can get the  $\ell_1$  error bound as:*

$$\|\widehat{B}_2 - \mathbb{B}\|_1 = O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/MR^2}\right)^{1/4}.$$

So far we have proved Theorem 1.2 for Algorithm 2.

#### 4.4 Refinement

For a given PSD matrix  $X = UU^\top$ , whose SVD is given by  $X = V\Sigma V^\top$ , we define  $X^{1/2}$  to be  $X^{1/2} = V\Sigma^{1/2}$ . Note that  $V = UH$  for some unknown rotation matrix  $H$ .

**Lemma 4.8** (Refinement with separation). *Recall that initialization  $\widehat{B}_1$  obtained from Step 3 such that  $\|\widehat{B}_1 - D_S \widetilde{B} D_S\| \leq (w_{\max} MR^2 / w_{\min} N)^{1/4}$ . Assume that  $\sigma_{\min}(D_S \widetilde{B} D_S) > (w_{\max}^2 MR^2 / w_{\min}^2 N)^{1/4}$ .*

*Let  $\widehat{V}$  denote the  $R$  leading left singular vectors of  $\widehat{B}_1$ . Regularize  $B$  from the 4-th batch of samples in the same way as in Step 3. Set*

$$Y = (\widehat{V}^\top D_S \widetilde{B} D_S \widehat{V})^{-1/2} (\widehat{V}^\top D_S \widetilde{B} D_S).$$

*Let  $\widehat{B}_3 = Y^\top Y$  and  $\widehat{B} = D_S^{-1} \widehat{B}_3 D_S^{-1}$ . We can bound the spectral distance by*

$$\|\widehat{B}_3 - D_S \widetilde{B} D_S\|_F = O\left(\sqrt{\frac{MR}{N}}\right), \quad (44)$$

$$\|\widehat{B}_4 - \mathbb{B}\|_1 = O\left(\sqrt{\frac{MR}{N}}\right). \quad (45)$$

## 5 Sample complexity lower bounds for estimation vs testing

### 5.1 Lower bound for estimating probabilities

We reduce the estimation problem to the community detection for a specific set of model parameters.

Consider the following topic model with equal mixing weights, i.e.  $w = w^c = 1/2$ . For some constant  $C_\Delta = \Omega(1)$ , the two word distributions are given by:

$$p = \left[ \frac{1+C_\Delta}{M}, \dots, \frac{1+C_\Delta}{M}, \frac{1-C_\Delta}{M}, \dots, \frac{1-C_\Delta}{M} \right],$$

$$q = \left[ \frac{1-C_\Delta}{M}, \dots, \frac{1-C_\Delta}{M}, \frac{1+C_\Delta}{M}, \dots, \frac{1+C_\Delta}{M} \right].$$

The expectation of the sum of samples is given by

$$\mathbb{E}[B_N] = N \frac{1}{2} (pp^\top + qq^\top) = \frac{N}{M^2} \begin{bmatrix} 1+C_\Delta^2 & 1-C_\Delta^2 \\ 1-C_\Delta^2 & 1+C_\Delta^2 \end{bmatrix}.$$

Note that the expected row sum is in the order of  $\Omega(\frac{N}{M})$ . When  $N$  is small, with high probability the entries of the empirical sum  $B_N$  only take value either 0 or 1, and  $B_N$  approximately corresponds to a SBM ( $G(M, a/M, b/M)$ ) with parameter  $a = \frac{N}{M}(1+C_\Delta^2)$  and  $b = \frac{N}{M}(1-C_\Delta^2)$ .

If the number of sample document is large enough for any algorithm to estimating the dictionary vector  $p$  and  $q$  up to  $\ell_1$  accuracy  $\epsilon$  for a small constant  $\epsilon$ , it can then be used to achieve partial recovery in the corresponding SBM, namely correctly classify a  $\gamma$  proportion of all the nodes for some constant  $\gamma = \frac{\epsilon}{C_\Delta}$ .

According to Zhang & Zhou [52], there is a universal constant  $C > 0$  such that if  $(a-b)^2/(a+b) < c \log(1/\gamma)$ , then there is no algorithm that can recover a  $\gamma$ -correct partition in expectation. This suggests that a necessary condition for us to learn the distributions is that

$$\frac{(2(N/M)C_\Delta^2)^2}{2(N/M)} \geq c \log(C_\Delta/\epsilon),$$

namely  $(N/M) \geq c \log(C_\Delta/\epsilon)/2C_\Delta^4$ . In the well separated regime, this means that the sample complexity is at least linear in the vocabulary size  $M$ .

Note that this lower bound is in a sense a worst case constructed with a particular distribution of  $p$  and  $q$ , and for other choices of  $p$  and  $q$  it is possible that the sample complexity can be much lower than that  $\Omega(M)$ .

## 5.2 Lower bound for testing property of HMMs

In this section, we prove an information theoretic lower bound for testing whether a sequence of observations consists of independent draws from  $Unif[M]$ , versus being a sequence of observations generated by a 2-state HMM with observation distributions supported on  $\{1, \dots, M\}$ . Such a lower bound will immediately yield a lower bound for estimating various properties of HMMs, including estimating the entropy rate, as a sequence of independent draws from  $Unif[M]$  has entropy rate  $\log(M)$ , whereas the 2-state HMMs we consider have an entropy rate that is an additive constant lower. We note that this HMM lower bound is significantly stronger than the analogous task of testing whether a matrix of probabilities has rank 1 versus rank 2. Such a task corresponds to only using the bi-gram counts extracted from the sequence of observations. It is conceivable that by leveraging longer sequences (i.e.  $k$ -grams for  $k > 2$ ), more information can be extracted about the instance. While this is the case, as our lower bound shows, even with such information,  $\Theta(M)$  observations are required to perform this test and distinguish these two cases.

**Theorem 5.1** (Theorem 1.8 restated). *Consider a sequence observations from a HMM with two hidden states  $\{s_p, s_q\}$ , emission distributions  $p, q$  supported on  $M$  elements, and probability  $t = \Omega(1)$  of transitioning from  $s_p$  to  $s_q$  and from  $s_q$  to  $s_p$ . For sufficiently large  $M$ , given a sequence of  $N$*

observations for  $N = o(M)$ , it is information theoretically impossible to distinguish the case that the two emission distributions are well separated, i.e.  $\|p - q\|_1 \geq 1/2$ , from the case that both  $p$  and  $q$  are uniform distribution over  $[M]$ , namely the HMM is degenerate of rank 1.

In order to derive a lower bound for the sample complexity, it suffices to show that given a sequence of  $N = o(M)$  consecutive observations, one can not distinguish whether it is generated by a random instance from a class of 2-state HMMs (Definition 1.5) with well-separated emission distribution  $p$  and  $q$ , or the sequence is simply  $N$  i.i.d. samples from the uniform distribution over  $\mathcal{M}$ , namely a degenerate HMM with  $p = q$ .

We shall focus on a class of well-separated HMMs parameterized as below: a symmetric transition matrix  $T = \begin{bmatrix} 1-t & t \\ t & 1-t \end{bmatrix}$ , where we set the transition probability to  $t = 1/4$ ; the initial state distribution is  $\pi_p = \pi_q = 1/2$  over the two states  $s_p$  and  $s_q$ ; the corresponding emission distribution  $p$  and  $q$  are uniform over two disjoint subsets of the vocabulary,  $\mathcal{A}$  and  $\mathcal{M} \setminus \mathcal{A}$ , separately. Moreover, we treat the set  $\mathcal{A}$  as a random variable, which can be any of the  $\binom{M}{M/2}$  subsets of the vocabulary of size  $M/2$ , chosen with equal probability  $1/\binom{M}{M/2}$ . Note that there is a one to one mapping between the set  $\mathcal{A}$  and an instance in the class of well-separated HMM.

Now consider a random sequence of  $N$  words  $G_1^N = [g_1, \dots, g_N] \in \mathcal{M}^N$ . If this sequence is generated by an instance of the 2-state HMM denoted by  $\mathcal{A}$ , the joint probability of  $(G_1^N, \mathcal{A})$  is given by:

$$\Pr_2(G_1^N, \mathcal{A}) = \Pr_2(G_1^N | \mathcal{A}) \Pr_2(\mathcal{A}) = \Pr_2(G_1^N | \mathcal{A}) \frac{1}{\binom{M}{M/2}} \quad (46)$$

Moreover, given  $\mathcal{A}$ , since the support of  $p$  and  $q$  are disjoint over  $\mathcal{A}$  and  $\mathcal{M} \setminus \mathcal{A}$  by our assumption, we can perfectly infer the sequence of hidden states  $S_1^N(G_1^N, \mathcal{A}) = [s_1, \dots, s_N] \in \{s_p, s_q\}^N$  simply by the rule  $s_i = s_p$  if  $g_i \in \mathcal{A}$  and  $s_i = s_q$  otherwise. Thus we have:

$$\Pr_2(G_1^N | \mathcal{A}) = \Pr_2(G_1^N, S_1^N | \mathcal{A}) = \frac{1/2}{M/2} \prod_{i=2}^N \frac{(1-t)\mathbf{1}[s_i = s_{i-1}] + t\mathbf{1}[s_i \neq s_{i-1}]}{M/2}. \quad (47)$$

On the other hand, if the sequence  $G_1^N$  is simply i.i.d. samples from the uniform distribution over  $\mathcal{M}$ , its probability is given by

$$\Pr_1(G_1^N) = \frac{1}{M^N}. \quad (48)$$

We further define a joint distribution rule  $\Pr_1(G_1^N, \mathcal{A})$  such that the marginal probability agrees with  $\Pr_1(G_1^N)$ . In particular, we define:

$$\Pr_1(G_1^N, \mathcal{A}) = \Pr_1(\mathcal{A} | G_1^N) \Pr_1(G_1^N) \equiv \frac{\Pr_2(G_1^N | \mathcal{A})}{\sum_{\mathcal{B} \in \binom{M}{M/2}} \Pr_2(G_1^N | \mathcal{B})} \Pr_1(G_1^N), \quad (49)$$

where we define the conditional probability  $\Pr_1(\mathcal{A} | G_1^N)$  using the properties of the 2-state HMM class.

The main idea of the proof of Theorem 5.1 is to show that if  $N = o(M)$ , the total variation distance between  $\Pr_1$  and  $\Pr_2$  vanishes to zero. It follows immediately from the connection between the error bound of hypothesis testing and total variation distance between two probability rules, that if  $TV(\Pr_1(G_1^N), \Pr_2(G_1^N))$  is too small we are not able to test which probability rule the random sequence  $G_1^N$  is generated according to.

The detailed proofs are provided in Appendix D.

As an immediate corollary of this theorem, it follows that many natural properties of HMMs cannot be estimated using a sublinear length sequence of observations:

**Corollary 5.2.** *For HMMs with 2 states and emission distributions supported on a domain of size at most  $M$ , to estimate the entropy rate up to an additive constant  $c \leq 1$  requires a sequence of  $\Omega(M)$  observations.*

## A Proofs for Rank 2 Algorithm Phase I

*Proof.* (to Lemma 3.3 (Estimate the word marginal probability  $\rho$ ))

We analyze how accurate the empirical average  $\hat{\rho}$  is. Note that under the assumption of Poisson number of samples, we have  $\hat{\rho}_i \sim \frac{1}{N} \text{Poi}(N\rho_i)$ , and  $\text{Var}(\hat{\rho}_i) = \frac{1}{N}\rho_i$ . Apply Markov inequality:

$$\Pr\left(\sum_{i=1}^M \left|\frac{\hat{\rho}_i - \rho_i}{\sqrt{\rho_i}}\right|^2 > t\right) \leq \frac{M}{tN},$$

thus probability at least  $1 - \delta$ , we can bound

$$\sum_{i=1}^M \left|\frac{\hat{\rho}_i - \rho_i}{\sqrt{\rho_i}}\right|^2 \leq \frac{M}{N\delta}. \quad (50)$$

Then apply Cauchy-Schwartz, we have

$$\sum_{i=1}^M |\hat{\rho}_i - \rho_i| \leq \left(\sum_{i=1}^M \sqrt{\rho_i}^2 \sum_{i=1}^M \left|\frac{\hat{\rho}_i - \rho_i}{\sqrt{\rho_i}}\right|^2\right)^{1/2} \leq \frac{1}{\sqrt{d_0\delta}}.$$

□

*Proof.* (to Lemma 3.4 (Concentration of marginal probabilities in the heaviest bin))

Fix constants  $C_1 = \frac{1}{2}$  and  $C_2 = 2$ , apply Corollary F.5 of Poisson tail (note that for word in the heaviest bin, we have  $N\rho_i > d_0 \log M$  to be a super constant), we show that  $\hat{\rho}_i$  concentrates well:

$$\Pr(C_1 N\rho_i < \text{Poi}(N\rho_i) < C_2 N\rho_i) \geq 1 - 4e^{-N\rho_i/2} \geq 1 - 4e^{-N \log M/(2M)}.$$

Note that the number of words in the heaviest bin is upper bounded by  $M_{\log} \leq \frac{1}{\min_{i \in \mathcal{I}_{\log}} \rho_i} \leq \frac{M}{\log(M)}$ . Take a union bound, we have that with high probability, all the estimates  $\hat{\rho}_i$ 's in the heaviest bin concentrate well:

$$\begin{aligned} \Pr(\forall i \in \mathcal{I}_{\log} : C_1 \rho_i < \hat{\rho}_i < C_2 \rho_i) &\geq 1 - \frac{M}{\log M} e^{-N \log M/(2M)} \\ &\geq 1 - 4e^{-N \log M/(2M) + \log M - \log \log M} \\ &\geq 1 - M^{-(d_0/2-1)} \\ &\geq 1 - M^{-1}, \end{aligned}$$

where recall that  $d_0 = N/M$  is a large constant. □

*Proof.* (to Lemma 3.5 (Estimate the dictionary separation restricted to the empirical heaviest bin))

(1) First, we claim that with high probability, no word from  $\mathcal{I}_k$  for  $k \leq \log(M) - e^2$  is placed in  $\hat{\mathcal{I}}_{\log M}$ . Namely all the words in  $\hat{\mathcal{I}}_{\log M}$  have true marginals at least  $\Omega(\frac{\log M}{M})$ . This is easy to show, by the Corollary F.5 of Poisson tail bound, each of the word from the much lighter bins is placed in  $\hat{\mathcal{I}}_{\log M}$  with probability less than  $2e^{-N \log M/M}$ . Take a union bound over all words with marginal at least  $1/M$ , we can bound the probability that any of the words being placed in  $\hat{\mathcal{I}}_{\log M}$  by  $2Me^{-d_0 \log M} = O(M^{-d_0+1})$ .

(2) The appropriate scaling with the diagonal matrix  $\text{diag}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}})^{-1/2}$  on both sides of the diagonal block is very important, which allows us to apply matrix Bernstein inequality at a sharper rate.

Note that with the two independent batches of samples, the empirical count matrix  $B$  considered here is independent from the empirical marginal vector  $\hat{\rho}$ . Thus for every fixed realization of  $\hat{\rho}$ , we have that with probability at least  $1 - M^{-1}$ ,

$$\begin{aligned}\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}(B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \mathbb{B}_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}})\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}\|_2 &\leq \sqrt{\frac{\log(M_{\log}/\delta)}{N}} + \frac{\frac{M}{\log(M)} \log(M_{\log}/\delta)}{N} \\ &= O\left(\frac{M}{N}\left(1 + \frac{\log(1/\delta)}{\log(M)}\right)\right) \\ &= O\left(\frac{M}{N}\right),\end{aligned}$$

where we used the fact that the all the marginals in the heaviest bin can be estimated with constant multiplicative accuracy given by Lemma 3.4; also, note that compared to the Bernstein matrix inequality directly applied to the entire matrix as in (72), here with the proper scaling we have  $\text{Var} \leq 1$  and  $B \leq \frac{M}{\log(M)}$ , since  $\hat{\rho}_i > \log(M)/M$  for all  $i \in \hat{\mathcal{I}}_{\log}$ .

We will show that  $B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}}$ , the diagonal block of the empirical count matrix, concentrates well enough to ensure that we can estimate the separation restricted to the heaviest bin by the leading eigenvector of  $(B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top)$ . Note that

$$\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \mathbb{B}_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} = \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \rho_{\hat{\mathcal{I}}_{\log}} (\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \rho_{\hat{\mathcal{I}}_{\log}})^\top + \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta'_{\hat{\mathcal{I}}_{\log}} (\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}})^\top.$$

Apply triangle inequality we have

$$\begin{aligned}&\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}(B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} - \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta'_{\hat{\mathcal{I}}_{\log}} (\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}})^\top\|_2 \\ &\leq \left\| \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}(B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \mathbb{B}_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}}) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \right\|_2 + \left\| \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}(\hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top - \rho_{\hat{\mathcal{I}}_{\log}} \rho_{\hat{\mathcal{I}}_{\log}}^\top) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \right\|_2 \\ &= O\left(\frac{M}{N}\right) + \sqrt{\frac{M}{N}} \\ &= O\left(\sqrt{\frac{M}{N}}\right).\end{aligned}$$

(3) Let  $uu^\top$  be the rank-1 truncated SVD of  $\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}(B_{\hat{\mathcal{I}}_{\log}, \hat{\mathcal{I}}_{\log}} - \hat{\rho}_{\hat{\mathcal{I}}_{\log}} \hat{\rho}_{\hat{\mathcal{I}}_{\log}}^\top) \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2}$ . Let  $v = \hat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} u$  be our estimate for  $\Delta_{\hat{\mathcal{I}}_{\log}}$ . Apply Wedin's theorem to rank-1 matrix (Lemma F.1), we can bound the distance between vector  $u$  and  $\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}}$  by:

$$\min\{\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}} - u\|_2, \|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}} + u\|_2\} = O\left(\min\left\{\frac{(M/N)^{1/2}}{\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} \Delta_{\hat{\mathcal{I}}_{\log}}\|_2}, (M/N)^{1/4}\right\}\right).$$

Note that  $\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} \mathbf{1}\|_2 = \|\hat{\rho}_{\hat{\mathcal{I}}_{\log}}^{1/2}\|_2 = 1$ . Apply Cauchy-Schwartz, for any vector  $x$ , we have

$$\|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} x\|_2 \geq \|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} x\|_2 \|\hat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} \mathbf{1}\|_2 \geq \left| \langle \hat{D}_{\hat{\mathcal{I}}_{\log}}^{-1/2} x, \hat{D}_{\hat{\mathcal{I}}_{\log}}^{1/2} \mathbf{1} \rangle \right| = \|x\|_1,$$



therefore, we can bound

$$\min\{\|\Delta_{\hat{\mathcal{I}}_{\log}} - v\|_1, \|\Delta_{\hat{\mathcal{I}}_{\log}} + v\|_1\} \leq O\left(\min\left\{\frac{(M/N)^{1/2}}{\|\Delta_{\hat{\mathcal{I}}_{\log}}\|_1}, (M/N)^{1/4}\right\}\right).$$

In the above inequalities we absorb all the universal constants and focus on the scaling factors.  $\square$

*Proof.* (to Lemma 3.6 (Spillover from much heavier bins is small in all bins))

Define  $d_k = e^k d_0$ , which is not to be confused with  $d_k^{\max}$  defined in (18).

(1) Consider  $k' = k + \tau$ . The probability that a word  $i$  from  $\mathcal{I}_{k'}$  falls into  $\hat{\mathcal{I}}_k$  is bounded by:

$$\Pr\left(N\frac{e^{k-1}}{M} < \text{Poi}(N\rho_i) < N\frac{e^k}{M}\right) < \Pr\left(\text{Poi}(N\frac{e^{(\tau+k)}}{M}) < N\frac{e^k}{M}\right) \leq 2e^{-e^{\tau+k}d_0/2} \quad (51)$$

where we apply the Poisson tail bound (1) in Corollary F.5, and set  $c = e^{-\tau} < 1/2$  for  $\tau \geq 1$ . Note that this bound is doubly exponentially decreasing in  $\tau$  and exponentially decreasing in  $d_0$ .

In expectation, we can bound  $\bar{W}_k$  by:

$$\begin{aligned} \mathbb{E}\bar{W}_k &= \mathbb{E} \sum_{i \in \mathcal{M}} \rho_i \sum_{k': k' \geq k + \tau + 1} \mathbf{1}[i \in \mathcal{I}_{k'}] \Pr\left(N\frac{e^{k-1}}{M} < \text{Poi}(N\rho_{k'}) < N\frac{e^k}{M}\right) \\ &\leq \sum_{i \in \mathcal{M}} \rho_i \Pr\left(\text{Poi}(N\frac{e^{(\tau+k)}}{M}) < N\frac{e^k}{M}\right) \\ &\leq 2e^{-e^{\tau+k}d_0/2}. \end{aligned}$$

Similarly, apply the Poisson tail bound (2) in Corollary F.5, and set  $c' = e^\tau \geq e$  for  $\tau \geq 1$ , we can bound the probability with which a word  $i$  from much lighter bins, namely  $\cup_{\{k': k' < k - \tau\}} \mathcal{I}_{k'}$ , is placed in the empirical bin  $\hat{\mathcal{I}}_k$  by:

$$\Pr\left(N\frac{e^k}{M} < \text{Poi}(N\rho_i) \leq N\frac{e^{k+1}}{M}\right) \leq \Pr\left(\text{Poi}(N\frac{e^{(k-\tau)}}{M}) > N\frac{e^k}{M}\right) \leq 2e^{-e^k d_0}, \quad (52)$$

and bound the total marginal probability by:

$$\mathbb{E}W_k \leq 2e^{-e^k d_0}.$$

(2) Next, we apply Bernstein's bound to get a high probability argument. We show that with high probability, for all the  $\log \log(M)$  bins, we can bound the spillover probability mass by  $\bar{W}_k \leq \mathbb{E}[\bar{W}_k] + O(\frac{1}{\text{poly}(M)})$ , which implies that asymptotically as the vocabulary size  $M \rightarrow \infty$ , we have  $\bar{W}_k \leq 2e^{-e^{\tau+k}d_0/2}$  for all  $k$ .

Consider the word  $i$  from the exact bin  $\mathcal{I}_{k'}$ , for some  $k' \geq k + \tau$ . Let

$$\lambda_i = 2e^{-e^{k'}d_0/2}$$

denote the upper bound (as shown in (51)) of the independent probability with which word  $i$  is placed in the empirical bin  $\hat{\mathcal{I}}_k$  (recall the Poisson number of samples assumption). The spillover probability mass is a random variable and can be written as

$$\bar{W}_k = \sum_{i \in \mathcal{I}_{k'}: (k+\tau) < k' \leq \log \log(M)} \rho_i \text{Ber}(\lambda_i),$$

Note that the summation of word  $i$  is over all the bin  $\mathcal{I}_{k'}$  for  $(k + \tau) \leq k' \leq \log \log(M)$ , where recall that in Lemma 3.4 we showed that with high probability the heaviest words are retained in the empirical bin  $\widehat{\mathcal{I}}_{\log}$ . Apply Bernstein's inequality to bound  $\overline{W}_k$ :

$$\Pr(\overline{W}_k - \mathbb{E}\overline{W}_k > t) \leq e^{-\frac{t^2/2}{\sum_i \rho_i^2 \lambda_i + \max_i \rho_i t/3}}.$$

To ensure that the right hand side is bounded by  $e^{-\log M}$  (this is to create space for the union bound over the  $\log \log M$  bins), we can fix some large universal constant  $C$  and set  $t$  to be

$$t = 2 \left( \left( \sum_i \rho_i^2 \lambda_i \right)^{1/2} + \max_i \rho_i \right) \log(M).$$

which right hand side can be bounded by:

$$\begin{aligned} \left( \sum_i \rho_i^2 \lambda_i \right)^{1/2} + \max_i \rho_i &\leq \left( \max_{i \in \mathcal{I}_{k'}: (k+\tau) \leq k' \leq \log \log(M)} \left( \frac{1}{\rho_i} \right) (\rho_i^2) (2e^{-e^{k'} d_0/2}) \right)^{1/2} + \frac{\log M}{M} \\ &\leq \left( 2 \max_{(k+\tau) \leq k' \leq \log \log(M)} \frac{e^{k'}}{M} e^{-e^{k'} d_0/2} \right)^{1/2} + \frac{\log M}{M} \\ &\leq \frac{2e^{-e^{k+\tau} d_0/4}}{\sqrt{M}} + \frac{\log M}{M}, \end{aligned}$$

where the first inequality is uses the worst case to bound the summation, and the last inequality uses the fact that  $d_0 = \Omega(1)$  is a large constant. Therefore, we can set  $t = 2 \left( \frac{e^{-e^{k+\tau} d_0/4} \log M}{\sqrt{M}} + \frac{(\log M)^2}{M} \right)$ . Finally, take a union bound over at most  $\log \log(M)$  moderate bins, we argue that with high probability (at least  $1 - O(1/M)$ ), for all the empirical moderate bins, we can bound the spillover marginal by:

$$\overline{W}_k \leq \mathbb{E}\overline{W}_k + O\left(\frac{1}{\text{poly}(M)}\right).$$

**(3)** Moreover, assume that  $W_k \geq e^{-k}$ , we can bound the number of the heavy spillover words  $\overline{M}_k$  compared to number of words in the exact bin  $M_k$ .

First note that  $\overline{M}_k \leq \frac{\overline{W}_k}{e^{\tau+k}/M}$ . Recall that  $d_k^{\max} = NW_k(e^{\tau+k}/M)$  was defined in (18). Also, since  $W_k \geq e^{-k} \gg \overline{W}_k \approx e^{-e^{k+\tau} d_0}$ , we can lower bound the number of words in the empirical bin  $\widehat{\mathcal{I}}_1$  by:

$$M_k \geq \frac{W_k}{e^{k+\tau}/M}.$$

Thus we can bound

$$\begin{aligned} \overline{M}_k \frac{d_k^{\max}}{M_k} &\leq \left( \frac{\overline{W}_k}{e^{k+\tau}/M} \right) \frac{(NW_k(e^{k+\tau}/M))}{\left( \frac{W_k}{e^{k+\tau}/M} \right)} \\ &\leq e^{k+\tau} d_0 \overline{W}_k \\ &\leq \frac{2e^{k+\tau} d_0}{e^{e^{k+\tau} d_0}} \\ &\leq 1, \end{aligned}$$

where the second last inequality we used the high probability upper bound for  $\overline{W}_k$ , and in the last inequality we use the fact that  $e^x > 2x$  for all  $x$ .  $\square$

*Proof.* (of Lemma 3.7 (Concentration of the regularized diagonal block  $\tilde{B}_k$ .)

In Figure 2, the rows and the columns of  $B_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k}$  are sorted according to the exact marginal probabilities of the words in ascending order. The rows and columns that are set to 0 by regularization are shaded.

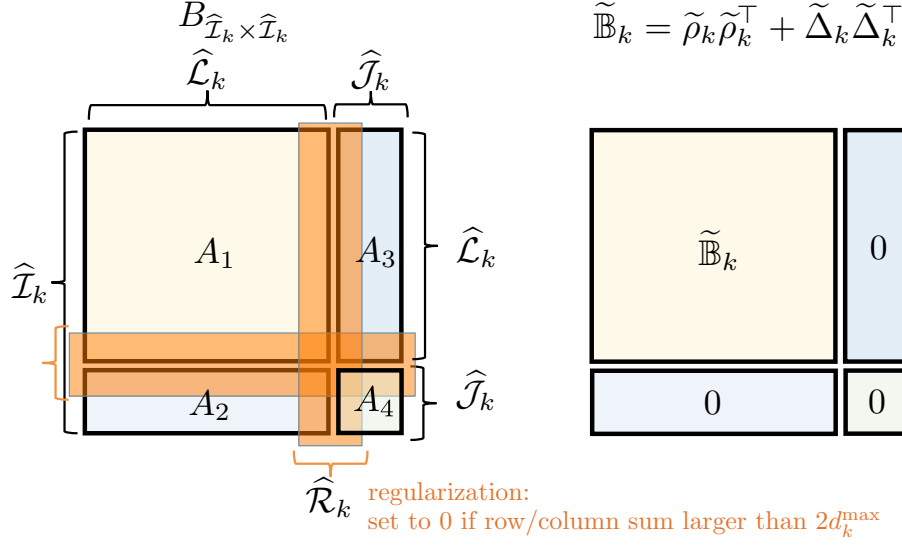


Figure 3: block decomposition of the diagonal block of  $B_{N2}$  corresponding to  $\hat{\mathcal{I}}_k$ .

On the left hand side, it is the empirical matrix without regularization. We denote the removed elements by matrix  $E \in \mathbb{R}_+^{M_k \times M_k}$ , whose only nonzero entries are those that are removed from in the regularization step (in the strips with orange color), namely  $E = [B_{i,j} \mathbf{1}[i \text{ or } j \in \hat{\mathcal{R}}_k]]$ . We denote the retained elements by matrix  $\tilde{B}_k = B_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k} \setminus E = B_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k} - E$ .

On the right hand side it is the same block decomposition applied to the matrix which we want the regularized empirical count matrix converges to. Recall that we defined  $\tilde{B}_k = \tilde{\rho}_k \tilde{\rho}_k^\top + \tilde{\Delta}_k \tilde{\Delta}_k^\top$  in (25), where we set entries corresponding to the words in the spillover set  $\hat{\mathcal{J}}_k$  to 0.

We bound the spectral distance of the 4 blocks  $(A_1, A_2, A_3, A_4)$  separately. The bound for the entire matrix  $\hat{B}_k$  is then an immediate result of triangle inequality:

$$\begin{aligned} \|\hat{B}_k - \tilde{B}_k\| &= \|[B_{N2}]_{\hat{\mathcal{I}}_k \times \hat{\mathcal{I}}_k} - E - \tilde{B}_k\| \\ &= \|A_1 \setminus E + A_2 \setminus E + A_3 \setminus E + A_4 \setminus E - N\tilde{B}_k\| \\ &\leq \|A_1 \setminus E - \mathbb{B}_{\hat{\mathcal{L}}_k \times \hat{\mathcal{L}}_k}\| + \|A_2 \setminus E\| + \|A_3 \setminus E\| + \|A_4 \setminus E\|. \end{aligned}$$

We bound the 4 parts separately below in **(a)-(c)**.

**(a)** To bound  $\|A_1 \setminus E - \mathbb{B}_{\hat{\mathcal{L}}_k \times \hat{\mathcal{L}}_k}\|$ , we first make a few observations:

1. By definition of  $\hat{\mathcal{J}}_k$  and  $\hat{\mathcal{L}}_k$ , every entry of the random matrix  $A_1$  is distributed as an independent Poisson variable  $\frac{1}{N} \text{Poi}(\lambda_k)$ , where  $\lambda_k \leq N(\frac{e^{k+\tau}}{M})^2 \leq d_0 \frac{\log(M)}{M} = o(1)$ .
2. The expected row sum of  $A_1$  is bounded by of  $d_k^{\max}$ .
3. With the regularization of removing the heavy rows and columns in  $E$ , every column sum and the row sum of  $A_1$  is bounded by  $2d_k^{\max}$ .

Therefore, by applying the Lemma F.6 (an immediate extension of the main theorem in [34]), we can argue that with probability at least  $1 - M_k^{-r}$  for some constant  $r = \Theta(1)$ ,

$$\|A_1 \setminus E - \mathbb{B}_{\hat{\mathcal{L}}_k \times \hat{\mathcal{L}}_k}\|_2 = O(\sqrt{d_k^{\max}/N}).$$

(b) To bound  $\|A_2 \setminus E\|$  and  $\|A_3 \setminus E\|$ , the key observations are:

1. Every row sum of  $A_2 \setminus E$  and every column sum of  $A_3 \setminus E$  is bounded by  $2d_k^{\max}$ .
2. For every non-zero row of  $A_2$ , its distribution is entry-wise dominated by a multinomial distribution  $\frac{1}{N} \text{Mul}\left(\frac{\rho_{\hat{\mathcal{L}}_k}}{\sum_{i \in \hat{\mathcal{L}}_k} \rho_i}; (2Nd_k^{\max})\right)$ , while the entries in  $E$  are set to 0, and note that in  $A_2$  the columns are restricted to the good words  $\hat{\mathcal{L}}_k$ . Moreover, by the Poisson assumption on  $N$  (recall that  $d_k^{\max} = W_k \bar{\rho}_k$ ), we have that the distributions of the entries in the row are independently dominated by  $\frac{1}{N} \text{Poi}\left(\frac{2Nd_k^{\max}}{M_k}\right)$ .

**Lemma A.1** (row/column-wise  $\ell_1$  norm to  $\ell_2$  norm bound (Lemma 2.5 in [34])). *Consider a matrix  $B$  in which each row has  $\mathcal{L}_1$  norm at most  $a$  and each column has  $\mathcal{L}_1$  norm at most  $b$ , then  $\|B\|_2 \leq \sqrt{ab}$ .*

**Claim A.2** (Sparse decomposition of  $(A_2 \setminus E)$ ). *With high probability, the index subset  $\hat{\mathcal{J}}_k \times \hat{\mathcal{L}}_k$  of  $(A_2 \setminus E)$  can be decomposed into two disjoint subsets  $\mathcal{R}$  and  $\mathcal{C}$  such that: each row of  $\mathcal{R}$  and each column of  $\mathcal{C}$  has row/column sum at most  $(\frac{r}{N} \log(Nd_k^{\max}))$ , for some constant  $r$ .*

Recall that from regularization we know that each column of  $\mathcal{R}$  and each row of  $\mathcal{C}$  in  $A_2 \setminus E$  has column/row sum at most  $2d_k^{\max}$ . Therefore we can apply Lemma A.1 and conclude that with high probability

$$\|A_2 \setminus E\|_2 \leq 2\sqrt{\frac{rd_k^{\max} \log(Nd_k^{\max})}{N}}.$$

*Proof.* (to Claim A.2)

We sketch the proof of Claim A.2, which mostly follows the sparse decomposition argument in Theorem 6.3 in [34]. We adapt their argument in our setup where the entries are distributed according to independent Poisson distributions. We first show (in (1)) that, with high probability, any square submatrix in  $(A_2 \setminus E)$  actually contains a sparse column with almost only a constant column sum; then, with this property we can (in (2)) iteratively take out sparse columns and rows from  $(A_2 \setminus E)$  to construct the  $\mathcal{R}$  and  $\mathcal{C}$ .

(1) With high probability, in any square submatrix of size  $m \times m$  in  $(A_2 \setminus E)$ , there exists a sparse column whose sum is at most  $(\frac{r}{N} \log(Nd_k^{\max}))$ .

To show this, consider an arbitrary column in an arbitrary submatrix of size  $m \times m$  in  $(A_2 \setminus E)$ . Recall our observation (b).2, that the column sum is dominated by  $\frac{1}{N} \text{Poi}(\lambda)$  with rate

$$\lambda = 2Nd_k^{\max} \frac{m}{M_k}.$$

Therefore, we can bound the column sum by applying the Chernoff bound for Poisson distribution

(Lemma F.2):

$$\begin{aligned}
\Pr\left(\text{a column sum} > \left(\frac{r}{N} \log Nd_k^{\max}\right)\right) &\leq \Pr(\text{Poi}(\lambda) > (r \log Nd_k^{\max})) \\
&\leq e^{-\lambda} \left(\frac{r \log Nd_k^{\max}}{e\lambda}\right)^{-r \log Nd_k^{\max}} \\
&\leq \left(\frac{rM_k}{2Nd_k^{\max}m}\right)^{-r \log Nd_k^{\max}} \\
&\leq \left(\frac{rM_k}{2m}\right)^{-r},
\end{aligned}$$

where in the last inequality we used the fact that for  $Nd_k^{\max}$  and  $r$  to be large constant, the following simple inequality holds:

$$\log(Nd_k^{\max}) \log\left(\frac{rM_k}{2Nd_k^{\max}m}\right) \geq \log\left(\frac{rM_k}{m}\right).$$

Then consider all the  $m$  columns in the submatrix of size  $m \times m$ , which column sums are independently dominated by Poisson distributions, we have

$$\Pr\left(\text{every column sum} > \left(\frac{r}{N} \log Nd_k^{\max}\right)\right) \leq \left(\frac{rM_k}{2m}\right)^{-rm}.$$

Next, take a union bound over all the  $m \times m$  submatrices of  $(A_2 \setminus E)$  for  $m$  ranging between 1 and  $\bar{M}_k$ , and recall that block  $(A_2 \setminus E)$  is of size  $\bar{M}_k \times (M_k - \bar{M}_k)$ . We can bound for all the submatrices:

$$\begin{aligned}
&\Pr\left(\text{for every submatrix in } (A_2 \setminus E), \text{ there exist a column whose sum} \leq \left(\frac{r}{N} \log Nd_k^{\max}\right)\right) \\
&\geq 1 - \sum_{m=1}^{\bar{M}_k} \binom{M_k}{m} \binom{\bar{M}_k}{m} \left(\frac{rM_k}{2m}\right)^{-rm} \\
&\geq 1 - \sum_{m=1}^{\bar{M}_k} \left(\frac{M_k}{m}\right)^{2m} \left(\frac{rM_k}{2m}\right)^{-rm} \\
&\geq 1 - M_k^{-(r-2)}.
\end{aligned} \tag{53}$$

Note that this is indeed a high probability event, since for  $W_k \geq \epsilon_0 e^{-k}$ , we have shown that  $M_k \geq M e^{-2k+\tau}$ .

**(2)** Perform iterative row and column deletion to construct  $\mathcal{R}$  and  $\mathcal{C}$ .

Given  $(A_2 \setminus E)$  of size  $\bar{M}_k \times M_k$ , we apply the argument above in (1) iteratively. First select a sparse column and remove it to  $\mathcal{C}$ , and apply it to remove columns until the remaining number of columns and rows are equal, then apply it alternatively to the rows (move to  $\mathcal{R}$ ) and columns (move to  $\mathcal{C}$ ) until empty. By construction, there are at most  $M_k$  such sparse columns in  $\mathcal{C}$ , each column of  $\mathcal{C}$  has sum bounded by  $(\frac{r}{N} \log Nd_k^{\max})$ , and each row of  $\mathcal{C}$  bounded by  $2d_k^{\max}$  because it is in the regularized  $(A_2 \setminus E)$ ; similarly  $\mathcal{R}$  has at most  $\bar{M}_k$  rows and each row of  $\mathcal{R}$  has sum at most  $(\frac{r}{N} \log Nd_k^{\max})$  and each column has sum at most  $2d_k^{\max}$ . □

The proof for the other narrow strip  $(A_3 \setminus E)$  is in parallel with the above analysis for  $(A_2 \setminus E)$ .

**(c)** To bound  $\|A_4 \setminus E\|$ , the two key observation are:

1. The total marginal probability mass of spillover heavy words  $\overline{W}_k = \sum_{i \in \hat{\mathcal{J}}_k} \rho_i \leq 2e^{-e^{k+\tau}d_0/2}$ . (shown in Lemma 4.1).
2. Similar to the observation in (b).2 above, the distributions of the entries in each row of  $(A_4 \setminus E)$  are independently dominated by  $\frac{1}{N} \text{Poi} \left( 2Nd_k^{\max} \frac{\overline{W}_k}{W_k} \frac{1}{\overline{M}_k} \right)$ .

In parallel with Claim A.2, we make a claim about the spectral norm of the block  $(A_4 \setminus E)$ :

**Claim A.3** (Sparse decomposition of  $(A_4 \setminus E)$ ). *With high probability, the index subset  $\hat{\mathcal{J}}_k \times \hat{\mathcal{J}}_k$  of  $A_2$  can be decomposed into two disjoint subsets  $\mathcal{R}$  and  $\mathcal{C}$  such that: each row of  $\mathcal{R}$  and each column of  $\mathcal{C}$  has sum at most  $\frac{r}{N}$ ; each column of  $\mathcal{R}$  and each row of  $\mathcal{C}$  has sum at most  $d_k^{\max}$ .*

*Proof.* (to Claim A.3)

To show this, we construct sparse decomposition similar to that of  $(A_2 \setminus E)$ .

The only difference is that, when considering all the  $m \times m$  submatrices, we only need to consider all the submatrices contained in the small square  $(A_4 \setminus E)$  of size  $\hat{\mathcal{J}}_k \times \hat{\mathcal{J}}_k$ , instead of all submatrices in the wide strip  $(A_2 \setminus E)$  of size  $\hat{\mathcal{L}}_k \times \hat{\mathcal{J}}_k$ . In this case, taking the union bound leads to factors of  $\overline{M}_k$ , compared to that of  $M_k$  in (53).

Here we only highlight the difference in the inequalities. Consider an arbitrary column in an arbitrary submatrix of size  $m \times m$  in  $(A_4 \setminus E)$ . Recall that this column sum is dominated by  $\frac{1}{N} \text{Poi}(\lambda)$  with rate

$$\lambda = 2Nd_k^{\max} \frac{\overline{W}_k}{W_k} \frac{m}{\overline{M}_k}.$$

Thus we can bound the probability of having a dense column by:

$$\Pr(\text{a column sum} > \frac{r}{N}) \leq \Pr(\text{Poi}(\lambda) > r) \leq e^{-\lambda} \left( \frac{r}{e\lambda} \right)^{-r} \leq \left( \frac{r}{e\lambda} \right)^{-r}.$$

Take a union over all the square matrices of size  $m \times m$  in block  $(A_4 \setminus E)$ , we can bound:

$$\begin{aligned} & \Pr(\text{for every submatrix in } (A_4 \setminus E), \text{ there exist a column whose sum} \leq \frac{r}{N}) \\ & \geq 1 - \sum_{m=1}^{\overline{M}_k} \binom{\overline{M}_k}{m} \left( \frac{r}{e\lambda} \right)^{-rm} \\ & \geq 1 - \sum_{m=1}^{\overline{M}_k} \left( \frac{\overline{M}_k}{m} \right)^{2m} \left( \frac{\overline{M}_k}{m} \frac{rW_k}{e2d_k^{\max}\overline{W}_k} \right)^{-rm} \\ & \geq 1 - M_k^{-(r-2)}, \end{aligned}$$

where in the last inequality we used the fact that  $d_k^{\max} = NW_k \frac{e^{k+\tau}}{M}$ , and plug in the high probability upper bound of  $\overline{W}_k \leq 2e^{-e^{k+\tau}d_0}$  as in (35), we have:

$$\frac{rW_k}{e2d_k^{\max}\overline{W}_k} = \frac{rW_k M}{2eNW_k e^{k+\tau} e^{-e^{k+\tau}d_0}} = \frac{re^{(e^{k+\tau}d_0)}}{2e(e^{k+\tau}d_0)} \gg 1.$$

Again note that given that the bin has significant total marginal probability, thus  $M_k \geq Me^{-2k}$ , the above probability bound is indeed a high probability statement.  $\square$

$\square$

*Proof.* (to Lemma 3.8 (Given spectral concentration of block  $\tilde{B}_k$ , estimate the separation  $\tilde{\Delta}_k$  ))

Recall the result of Lemma 3.7 about the concentration of the diagonal block with regularization. For empirical bin with large enough marginal  $W_k$ , we have with high probability,

$$\|\tilde{B}_k - \tilde{\mathbb{B}}_k\|_2 \leq C \sqrt{\frac{d_k^{\max} \log^2 d_k^{\max}}{N}}.$$

Also recall that  $\hat{\rho}_{\hat{\mathcal{I}}_k}$  is defined to be the exact marginal vector restricted to the empirical bin  $\hat{\mathcal{I}}_k$ .

We can also bound

$$\begin{aligned} \left\| (\tilde{B}_k - \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top) - \tilde{\Delta}_k \tilde{\Delta}_k^\top \right\|_2 &\leq \left\| (\tilde{B}_k - \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top) - (\tilde{\mathbb{B}}_k - \tilde{\rho}_k \tilde{\rho}_k^\top) \right\|_2 \\ &\leq \left\| \tilde{B}_k - \tilde{\mathbb{B}}_k \right\|_2 + \left\| \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top - \tilde{\rho}_k \tilde{\rho}_k^\top \right\|_2 \\ &\leq C \sqrt{\frac{d_k^{\max} \log^2 d_k^{\max}}{N}}. \end{aligned}$$

Note that in the last inequality above we ignored the term  $\|\hat{\rho}_{\hat{\mathcal{I}}_k} - \tilde{\rho}_k\|_2$  as it is small for all bins (with large probability):

$$\begin{aligned} \left\| \hat{\rho}_{\hat{\mathcal{I}}_k} \hat{\rho}_{\hat{\mathcal{I}}_k}^\top - \tilde{\rho}_k \tilde{\rho}_k^\top \right\|_2 &\leq 4 \left\| \hat{\rho}_{\hat{\mathcal{I}}_k} - \tilde{\rho}_k \right\|_2 \left\| \hat{\rho}_{\hat{\mathcal{I}}_k} \right\|_2 \leq \sqrt{\underbrace{\bar{\rho}_k (M_k/N)}_{\text{over } \hat{\mathcal{L}}_k} + \underbrace{\bar{\rho}_k^2 (\bar{W}_k / \bar{\rho}_k)}_{\text{over } \hat{\mathcal{J}}_k}} \sqrt{M_k \bar{\rho}_k^2} \\ &\leq \sqrt{\frac{M_k^2 \bar{\rho}_k^3}{N}} = o\left(\sqrt{\frac{d_k^{\max}}{N}}\right), \end{aligned}$$

where in the second inequality we write  $\left\| \hat{\rho}_{\hat{\mathcal{I}}_k} - \tilde{\rho}_k \right\|_2^2$  into two parts over the set of good words  $\hat{\mathcal{L}}_k$  and the set of bad words  $\hat{\mathcal{J}}_k$ . To bound the sum over  $\hat{\mathcal{L}}_k$  we used the Markov inequality as in the proof of Lemma 3.3; and to bound the sum over  $\hat{\mathcal{J}}_k$  as well as the term  $\left\| \hat{\rho}_{\hat{\mathcal{I}}_k} \right\|_2^2$  we used the fact that if a word  $i$  appears in  $\hat{\mathcal{I}}_k$ , we must have  $\hat{\rho}_i \leq \bar{\rho}_k$ . The last inequality is due to  $M_k^2 \bar{\rho}_k^3 \leq W_k^2 \bar{\rho}_k \leq W_k \bar{\rho}_k = d_k^{\max}$ .

Let  $v_k v_k^\top$  be the rank-1 truncated SVD of the regularized block  $(\tilde{B}_k - \hat{\rho}_k \hat{\rho}_k^\top)$ . Apply Wedin's theorem to rank-1 matrix (Lemma F.1), we can bound the distance between vector  $v_k$  and  $\tilde{\Delta}_k$  by:

$$\min \left\{ \|\tilde{\Delta}_k - v_k\|, \|\tilde{\Delta}_k + v_k\| \right\} = O \left( \min \left\{ \sqrt{\frac{d_k^{\max}}{N} \log(N d_k^{\max})} \frac{1}{\|\tilde{\Delta}_{\hat{\mathcal{I}}_k}\|_2}, \left( \sqrt{\frac{d_k^{\max}}{N} \log(N d_k^{\max})} \right)^{1/2} \right\} \right).$$

□

*Proof.* (to Lemma 3.11 (Accuracy of  $\hat{\Delta}$  in Phase I)) Consider for each empirical bin. If  $\widehat{W}_k < \epsilon_0 e^{-k}$  set  $\hat{\Delta}_{\hat{\mathcal{I}}_k} = 0$ . We can bound the total  $\ell_1$  norm error incurred in those bins by  $\epsilon_0$ . Also, for the lightest bin, we can bound the total  $\ell_1$  norm error from setting  $\hat{\Delta}_{\hat{\mathcal{I}}_0} = 0$  by  $\epsilon_0$  small constant. If  $\widehat{W}_k > \epsilon_0 e^{-k}$ , we can apply the concentration bounds in Lemma 3.4, 3.8, and note that  $\|\hat{\Delta}_{\hat{\mathcal{I}}_k} - \Delta_{\hat{\mathcal{I}}_k}\|_1 \leq \sqrt{M_k} \|\hat{\Delta}_{\hat{\mathcal{I}}_k} - \hat{\Delta}_{\hat{\mathcal{I}}_k}\|_2$ .

Note that we need to take a union bound of probability that spectral concentration results holds (Lemma 3.7) for all the bins with large enough marginal. This is true because we have at most  $\log \log M$  bins, and each bin's spectral concentration holds with high probability  $(1 - 1/\text{poly}(M))$ , thus even after taking the union bound the failure probability is still inverse poly in  $M$ .

Actually throughout the paper the small constant failure probability is only incurred when bounding the estimation error of  $\hat{\rho}$ , for the same reason of estimating a simple and unstructured distribution.

Overall, we can bound the estimation error in  $\ell_1$  norm by:

$$\begin{aligned}
\|\hat{\Delta} - \Delta\|_1 &\leq \underbrace{\epsilon_0}_{\text{lightest bin}} + \underbrace{1/d_0^{1/4}}_{\text{heaviest bin}} + \underbrace{\epsilon_0}_{\text{moderate bins with small marginal}} + \underbrace{\sum_k \sqrt{M_k} \left( \sqrt{\frac{d_k^{\max} \log N d_k^{\max}}{N}} \right)^{1/2}}_{\text{moderate bins with large marginal}} \\
&\leq 2\epsilon_0 + 1/d_0^{1/4} + \sum_k \left( \frac{M_k^2 W_k \bar{\rho}_k \log(N W_k \bar{\rho}_k)}{N} \right)^{1/4} \\
&\leq 2\epsilon_0 + (\log(d_0)/d_0)^{1/4} (1 + e^\tau \sum_k \left( \frac{W_k^2 M_k}{M} \right)^{1/4}) \\
&\leq 2\epsilon + (\log(d_0)/d_0)^{1/4} (1 + e^\tau) \\
&= O(\epsilon_0)
\end{aligned}$$

where in the second last inequality used Cauchy-Schwartz and the fact  $W_k \leq 1$ , so that  $\sum_k (W_k^2 M_k)^{1/4} \leq \sum_k (\sqrt{W_k} \sqrt{M_k})^{1/2} \leq (\sum_k W_k \sum_k M_k)^{1/4} \leq M^{1/4}$ , and in the last inequality above we use the assumption that  $d_0 =: N/M$  satisfies that  $d_0/\log(d_0) \geq 1/\epsilon_0^4$ .  $\square$

## B Proofs for Rank 2 Algorithm Phase II

*Proof.* (to Lemma 3.13 (Sufficient condition for constructing an anchor partition))

(1) First, we show that if for some constant  $c = \Omega(1)$ , a set of words  $\mathcal{A}$  satisfy

$$\left| \sum_{i \in \mathcal{A}} \Delta_i \right| \geq c \|\Delta\|_1, \tag{54}$$

then  $(\mathcal{A}, [M] \setminus \mathcal{A})$  is a pair of anchor set defined in 3.12.

By the assumption of constant separation ( $C_\Delta = \Omega(1)$ ),  $\sum_{i \in \mathcal{A}} \Delta_i = \Omega(1)$ . We can bound the condition number of the anchor partition matrix by:

$$\text{cond} \left( \begin{bmatrix} \rho_{\mathcal{A}} & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}} & -\Delta_{\mathcal{A}} \end{bmatrix} \right) = \frac{\sqrt{T^2 - 4D} + T}{\sqrt{T^2 - 4D} - T} \leq \frac{\sqrt{1 + 4cC_\Delta} + 1}{\sqrt{1 + 4cC_\Delta} - 1} = \Omega(1),$$

where  $T = \rho_{\mathcal{A}} - \Delta_{\mathcal{A}} \leq 1$  and  $D = -\rho_{\mathcal{A}} \Delta_{\mathcal{A}} - (1 - \rho_{\mathcal{A}}) \Delta_{\mathcal{A}} = -\Delta_{\mathcal{A}}$ .

(2) Next we show that  $\hat{\mathcal{A}}$  defined in the lemma statement satisfies (54).

Denote  $\mathcal{A}^* = \{i \in \mathcal{I} : \Delta_i > 0\}$ . Note that  $\|\Delta_{\mathcal{I}}\|_1 = \sum_{i \in \mathcal{A}^*} \Delta_i - \sum_{i \in \mathcal{I} \setminus \mathcal{A}^*} \Delta_i$ . Without loss of generality we assume that  $\sum_{i \in \mathcal{A}^*} \Delta_i \geq \frac{1}{2} \|\Delta_{\mathcal{I}}\|_1 \geq \frac{1}{2} C \|\Delta\|_1$ , where the last inequality is by the condition  $\|\Delta_{\mathcal{I}}\|_1 \geq C \|\Delta\|_1$ .



Given  $\hat{\Delta}_{\mathcal{I}}$  that satisfies (29). We look at  $\hat{\mathcal{A}} = \{i \in \mathcal{I} : \hat{\Delta}_i > 0\}$ .

$$\begin{aligned}
\sum_{i \in \hat{\mathcal{A}}} \Delta_i &= \sum_{i \in \hat{\mathcal{A}} \cap \mathcal{A}^*} \Delta_i - \sum_{i \in \hat{\mathcal{A}} \cap (\mathcal{I} \setminus \mathcal{A}^*)} \Delta_i \\
&= \sum_{i \in \mathcal{A}^*} \Delta_i - \sum_{i \in (\hat{\mathcal{A}} \cap (\mathcal{I} \setminus \mathcal{A}^*)) \cup (\mathcal{A}^* \cap (\mathcal{I} \setminus \hat{\mathcal{A}}))} |\Delta_i| \\
&\geq \sum_{i \in \mathcal{A}^*} \Delta_i - \|\hat{\Delta}_{\mathcal{I}} - \Delta_{\mathcal{I}}\|_1 \\
&\geq \left(\frac{1}{2}C - C'\right) \|\Delta\|_1 \\
&\geq \frac{1}{6}CC_{\Delta},
\end{aligned}$$

where in the second last inequality we used the fact that, if the sign of  $\hat{\Delta}_i$  and  $\Delta_i$  are different, it must be that  $|\hat{\Delta}_i - \Delta_i| > |\Delta_i|$ .  $\square$

*Proof.* (to Lemma 3.16 (Estimate the separation restricted to the  $k$ -th good bin))

Since it is a good bin, we have the  $\ell_2$  bound given by Lemma 3.8 as below (assuming the possible sign flip has been fixed as in Lemma 3.10):

$$\|\tilde{\Delta}_k - v_k\|_2 \leq \frac{\sqrt{d_k^{\max}} \log^2 d_k^{\max}}{N} \frac{1}{\|\tilde{\Delta}_k\|_2}.$$

Then we can convert the bound to  $\ell_1$  distance by:

$$\begin{aligned}
\frac{\|v_k - \tilde{\Delta}_k\|_1}{\|\tilde{\Delta}_k\|_1} &\leq \frac{\sqrt{M_k} \|v_k - \tilde{\Delta}_k\|_2}{\|\tilde{\Delta}_k\|_1} \leq \frac{\sqrt{M_k} \|v_k - \tilde{\Delta}_k\|_2 \|\tilde{\Delta}_k\|_2}{\|\tilde{\Delta}_k\|_2 \|\tilde{\Delta}_k\|_1} \\
&\leq \frac{M_k \|v_k v_k^\top - \tilde{\Delta}_k \tilde{\Delta}_k^\top\|}{\|\tilde{\Delta}_k\|_1^2} \leq C \frac{M_k}{W_k^2} \sqrt{d_k^{\max}} \frac{\log^2(d_0)}{N} \\
&\leq C \frac{M_k}{W_k^2} e^\tau \sqrt{N W_k \frac{W_k}{M_k} \frac{\log^2(d_0)}{N}} \\
&\leq C e^\tau \sqrt{\frac{M \log^2(d_0)}{N W_k e^k}} = O\left(\sqrt{\frac{\log(d_0)}{d_0}}\right),
\end{aligned}$$

where in the second last inequality, we used the fact that  $M_k \frac{e^k}{M} \leq W_k$  again, and in the last inequality we used the assumption  $W_k \geq \epsilon_0/e^k$ .  $\square$

*Proof.* (to Lemma 3.15 (With  $\Omega(1)$  separation, most words fall in “good bins” with high probability))

This proof is mostly playing around with the probability mass and converting something obviously true in expectation to high probability argument.

(1) Note that by their definition we know that  $W_k \geq \frac{1}{2}S_k$ , and we have

$$\begin{aligned}
& \sum_k W_k \left( \mathbf{1}[\frac{S_k}{2W_k} \geq C_2] + \frac{S_k}{2W_k} \mathbf{1}[\frac{S_k}{2W_k} < C_2] \right) \\
& \geq \sum_k W_k \frac{S_k}{2W_k} \left( \mathbf{1}[\frac{S_k}{2W_k} \geq C_2] + \mathbf{1}[\frac{S_k}{2W_k} < C_2] \right) \\
& = \sum_k W_k \frac{S_k}{2W_k} \\
& = \frac{1}{2} \sum_k S_k,
\end{aligned}$$

Moreover, note that by definition of  $W_k$ , we have  $\sum_k W_k = 1$ , therefore

$$\sum_k W_k \frac{S_k}{2W_k} \mathbf{1}[\frac{S_k}{2W_k} < C_2] \leq \sum_k C_2 W_k = C_2.$$

From the above two inequalities we can bound

$$\sum_k W_k \mathbf{1}[\frac{S_k}{2W_k} \geq C_2] \geq \frac{1}{2} \sum_k S_k - C_2.$$

Also note that

$$\sum_k W_k \mathbf{1}[W_k < \frac{C_1}{2^k}] \leq C_1.$$

Therefore according to the definition of “good bins” we have that:

$$\sum_{k \in \mathcal{G}} W_k = \sum_k W_k \mathbf{1} \left[ \frac{S_k}{2W_k} \geq C_2 \text{ and } W_k \geq \frac{C_1}{2^k} \right] \geq \frac{1}{2} \sum_k S_k - C_2 - C_1. \quad (55)$$

(2) We want to lower bound the quantity  $\sum_k S_k$  to be a constant fraction of  $\|\Delta\|_1$ . Note that by definition of  $S_k$  we can equivalently write the sum as:

$$\sum_k S_k = \sum_k \sum_{i \in \widehat{\mathcal{I}}_k \cap (\cup_{\{k':k' \leq k+\tau\}} \mathcal{I}_{k'})} |\Delta_i| = \sum_i |\Delta_i| \sum_k \mathbf{1}[i \in \mathcal{I}_k \cap \cup_{\{k':k' \leq k+\tau\}} \widehat{\mathcal{I}}_{k'}].$$

Consider for each word  $i$ . Assume that word  $i \in \mathcal{I}_k$ . Given  $N = d_0 M$  for some large constant  $d_0$ , denote  $d_k = e^k d_0$ , we can bound the probability  $\Pr(i \in \mathcal{I}_k \cap \cup_{\{k':k' \leq k+\tau\}} \widehat{\mathcal{I}}_{k'})$  as follows:

$$\begin{aligned}
\Pr(i \in \mathcal{I}_k \cap \cup_{\{k':k' \leq k+\tau\}} \widehat{\mathcal{I}}_{k'}) & \geq 1 - \Pr(\text{Poi}(N\rho_i) > e^\tau \rho_i) - \Pr(\text{Poi}(N\rho_i) < e^{-\tau} N\rho_i/2) \\
& \geq 1 - \frac{e^{-(\tau-1)e^{(\tau+k)}d_0}}{\sqrt{2\pi}e^{k+\tau}d_0} - e^{-d_k} \\
& \geq 1 - 2e^{-d_k}.
\end{aligned}$$

Therefore at least in expectation we can lower bound the sum by

$$\mathbb{E}[\sum_k S_k] = \sum_i |\Delta_i| \sum_k \Pr(i \in \mathcal{I}_k \cap \cup_{\{k':k' \leq k+\tau\}} \widehat{\mathcal{I}}_{k'}) \geq (1 - 2e^{-d_0}) \|\Delta\|_1.$$

(3) Restrict to the exact good bins, for which we know that the exact  $\|\rho_{\hat{\mathcal{I}}_k}\|_1 \geq e^{-k}$  and  $\|\Delta_{\hat{\mathcal{I}}_k}\|_1/\|\rho_{\mathcal{I}_k}\|_1 \geq C$ .

Here we know that if  $\hat{\mathcal{I}}_k$  is an good bin, the number of words in this exact good bin is lower bounded by  $M_k \geq e^{-k}/\rho_k \geq M/e^{-k}$ , and since  $k \leq \log \log M$  we have that  $M_k \geq \frac{M}{\log(M)}$ . This is important for use to apply Bernstein concentration of the words in the bin.

Since  $\|\Delta_{\hat{\mathcal{I}}_k}\|_1/\|\rho_{\hat{\mathcal{I}}_k}\|_1 \geq C$ , and that  $|\Delta_i| \leq \rho_i$ , we have that out of the  $M_k$  words there are at least a constant fraction of words with  $|\Delta_i| \geq \frac{1}{2}C\rho_k$ . Recall that we denote  $\rho_k = e^k/M$ . This is easy to see as  $x\rho_k + (M_k - x)\frac{1}{2}C\rho_k \geq \|\Delta_{\hat{\mathcal{I}}_k}\|_1 \geq CM_k\rho_k$  thus  $x \geq C/2 - CM_k$ .

Then we bound the probability that out of these  $cM_k$  words with good separation, a constant fraction of them do not escape from the closest  $\tau$  empirical bins. Denote  $\lambda_k = 2e^{-d_k}$ , which is the upper bound of the escaping probability for each of the word, and is very small. By a simple application of Bernstein bounds of the Bernoulli sum, for a small constant  $c_0$ , we have

$$\begin{aligned} \Pr\left(\sum_{i=1, \dots, cM_k} \text{Ber}_i(\lambda_k) \geq c_0M_k\right) &\leq \Pr\left(\sum_{i=1, \dots, cM_k} \text{Ber}_i(\lambda_k) - \lambda_kM_k \geq (c_0 - \lambda_k)M_k\right) \\ &\leq e^{-\frac{\frac{1}{2}(c_0 - \lambda_k)^2 M_k^2}{M_k \lambda_k + \frac{1}{3}(c_0 - \lambda_k)M_k}} \\ &\approx e^{-c_0M_k}. \end{aligned}$$

Then union bound over all the exact good bins. That gives a  $\log \log M$  multiply of the probability.

We now know that restricting to the non-escaping good words in the exact good bins, they already contribute a constant fraction (due to constant non-escaping, constant ratio  $\|\Delta_{\hat{\mathcal{I}}_k}\|_1/\|\rho_{\hat{\mathcal{I}}_k}\|_1$ , and constant  $\sum_{k \in \text{exact good bin}} W_k$ ) of the total separation  $\|\Delta\|_1$ . Therefore we can conclude that for some universal constant  $C$  we have

$$\sum_k S_k \geq C\|\Delta\|_1.$$

(4) Finally plug the above bound of  $\sum_k S_k$  into (55), and note the assumption on the constants  $C_1$  and  $C_2$ , we can conclude that the total marginal probability mass contained in “good bins” is large:

$$\sum_{k \in \mathcal{G}} W_k \geq (C - \frac{1}{24} - \frac{1}{24})\|\Delta\|_1 = \frac{1}{12}\|\Delta\|_1.$$

□

*Proof.* (to Lemma 3.17 (Estimate  $\rho$  and  $\Delta$  with accuracy in  $\ell_2$  distance))

Consider word  $i$  we have that

$$\begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix} \begin{bmatrix} \rho_i \\ \Delta_i \end{bmatrix} = \begin{bmatrix} \sum_{j \in \mathcal{A}} \mathbb{B}_{j,i} \\ \sum_{j \in \mathcal{A}^c} \mathbb{B}_{j,i} \end{bmatrix}$$

Set

$$\begin{bmatrix} \hat{\rho}_i \\ \hat{\Delta}_i \end{bmatrix} = \begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1 - \rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j \in \mathcal{A}} B_{j,i} \\ \sum_{j \in \mathcal{A}^c} B_{j,i} \end{bmatrix}$$

Since  $\sum_{j \in \mathcal{A}} B_{j,i} \sim \frac{1}{N} \text{Poi}(N(\rho_{\mathcal{A}}\rho_i + \Delta_{\mathcal{A}}\Delta_i))$ , apply Markov inequality, we have that

$$\Pr\left(\sum_i \left(\sum_{j \in \mathcal{A}} B_{j,i} - \sum_{j \in \mathcal{A}} \mathbb{B}_{j,i}\right)^2 > \epsilon^2\right) \leq \frac{\frac{1}{N} \sum_i (\rho_{\mathcal{A}}\rho_i + \Delta_{\mathcal{A}}\Delta_i)}{\epsilon^2} = \frac{\rho_{\mathcal{A}}}{N\epsilon^2}$$

Note that  $\rho_{\mathcal{A}} = \Omega(1)$  and that  $\text{cond}(D_{\mathcal{A}}) = \Omega(1)$ , we can propagate the concentration to the estimation error of  $\rho$  and  $\Delta$  as, for some constant  $C = \Omega(1)$ ,

$$\Pr(\|\hat{\rho} - \rho\| > \epsilon) \leq \frac{C}{N\epsilon^2}, \quad \Pr(\|\hat{\Delta} - \Delta\| > \epsilon) \leq \frac{C}{N\epsilon^2}.$$

□

## C Proofs for Rank $R$ Algorithm

*Proof.* (to Lemma 4.1 )

$$\mathbb{E}[\overline{W}_k] \leq 2e^{-e^{\tau+k}d_0/2}.$$

First we argue that with high probability, all words in bins  $k > \log \log M$  concentrates well. For  $\rho_i > \frac{\log M}{M}$ , set constant  $C = 1/2$ , we have

$$\Pr(\text{Poi}(N\rho_i) < \frac{1}{2}N\rho_i) \leq 2e^{-N\rho_i/2} \leq 2e^{N \log M/2M}.$$

There are at most  $\frac{M}{\log M}$  such heavy words. Take a union bound over them we have

$$\begin{aligned} \Pr(\forall i \text{ s.t. } i \in \mathcal{I}_k, k > \log \log M : i \in \widehat{\mathcal{I}}_{k'}, k' < k-1) &\geq 1 - \frac{M}{\log M} 2e^{-N \log M/2M} \\ &\geq 1 - 2\exp(-N \log M/2M + \log M - \log \log M) \\ &\geq 1 - 2\exp(-N \log M/4M) \\ &= 1 - 2M^{-d_0/4}. \end{aligned}$$

Second, define  $\overline{\mathcal{I}}_k = \{i : i \in \mathcal{I}_{k'}, k + \tau < k' < \log \log M\}$ . For word  $i \in \mathcal{I}_{k'}$ , let  $\lambda_i = 2e^{-e^{k'}d_0/2}$ , we have  $\overline{W}_k = \sum_{i \in \overline{\mathcal{I}}_k} \rho_i \text{Ber}(\lambda_i)$ . By Bernstein inequality:

$$\Pr(\overline{W}_k - \mathbb{E}\overline{W}_k > t) \leq \exp\left(-\frac{t^2}{\sum_{i \in \overline{\mathcal{I}}_k} \rho_i^2 \lambda_i + \max_{i \in \overline{\mathcal{I}}_k} \rho_i t}\right).$$

In order to bound the probability by  $\exp(-2 \log \log M)$ , so that we can take a union bound over the  $\log M$  bins, we set  $t$  to be  $t = 2 \log \log M (1/\sqrt{M} + \log M/M) = O(1/\text{poly}(M))$ , and note that

$$\begin{aligned} \left(\left(\sum_{i \in \overline{\mathcal{I}}_k} \rho_i^2 \lambda_i\right)^{1/2} + \max_{i \in \overline{\mathcal{I}}_k} \rho_i\right) &\leq \left(\max_{i \in \overline{\mathcal{I}}_k} \frac{1}{\rho_i} \rho_i^2 \lambda_i\right)^{1/2} + \frac{\log M}{M} \\ &\leq \left(\max_{k' > k+\tau} (e^{k'}/M) 2e^{-e^{k'}d_0/2}\right)^{1/2} + \log M/M \\ &\leq 1/\sqrt{M} + \log M/M. \end{aligned}$$

Therefore, we argue that with high probability, for all empirical bins  $\widehat{\mathcal{I}}_k$ , we can bound the spillover probability from heavy bins by:

$$\overline{W}_k \leq e^{-e^{\tau+k}d_0/2}.$$

□

*Proof.* (to Lemma 4.3 ) Consider for a typical word in the bin  $\mathcal{I}_k$ , we can bound the probability that it is not contained in bin  $\widehat{\mathcal{I}}_k$  by:

$$\Pr(\text{Poi}(N\rho_i) < \frac{1}{2}N\rho_i \text{ or } \text{Poi}(N\rho_i) > 2N\rho_i) \leq 4e^{-e^k d_0/2}.$$

Apply Bernstein inequality to all the  $W_k/\bar{\rho}_k$  words in bin  $\mathcal{I}_k$ , denote  $\lambda_k = 4e^{-e^k d_0/2}$ , we have

$$\Pr(W_k^s - \mathbb{E}W_k^s > t) < \exp\left(-\frac{t^2}{M_k \bar{\rho}_k^2 \lambda_k + \bar{\rho}_k t}\right)$$

Since the bin is big, we have  $W_k > e^{-k}$ , we can set

$$t = (M_k \bar{\rho}_k (\lambda_k/M_k)^{1/2} + \bar{\rho}_k) \log \log M \leq 4W_k e^{-e^k d_0/4} \frac{\log M}{\sqrt{M_k}} + \frac{\log M \log \log M}{M} = o(1),$$

where the last inequality is due to  $M_k > Me^{-2k}$ . Take a union bound over all  $\log M$  bins, we can ensure that with high probability, for each bin, the escaped mass is bounded by  $4W_k e^{-e^k d_0/2}$ .  $\square$

*Proof.* (to Lemma 4.5 ) In parallel with the analysis for Rank 2 (see Lemma 3.7), we know that regularization restores spectral concentration in the diagonal blocks. Denote the noise matrix in the regularized diagonal block by  $E_k = \widetilde{B}_k - \widetilde{\mathbb{B}}_k$ .

$$\|E_k\| = \|\widetilde{B}_k - \widetilde{\mathbb{B}}_k\| = O\left(\frac{\sqrt{Nd_k^{max} \log Nd_k^{max}}}{N}\right)$$

Denote the  $R$ -SVD of  $\widetilde{B}_k$  by  $\widehat{V}_k \widehat{\Lambda}_k \widehat{V}_k^\top$ .

$$\begin{aligned} \|\text{Proj}_{\widehat{V}_k} \widetilde{\mathbb{B}}_k \text{Proj}_{\widehat{V}_k} - \widetilde{\mathbb{B}}_k\| &= \|\text{Proj}_{\widehat{V}_k} (\widetilde{B}_k - E_k) \text{Proj}_{\widehat{V}_k} - \widetilde{\mathbb{B}}_k\| \\ &\stackrel{(a)}{\leq} \|\text{Proj}_{\widehat{V}_k} \widetilde{B}_k \text{Proj}_{\widehat{V}_k} - \widetilde{\mathbb{B}}_k\| + \|\text{Proj}_{\widehat{V}_k} E_k\| \\ &\stackrel{(b)}{\leq} \|\text{Proj}_{\widehat{V}_k} \widetilde{B}_k \text{Proj}_{\widehat{V}_k} - \widetilde{B}_k\| + \|\widetilde{B}_k - \widetilde{\mathbb{B}}_k\| + \|\text{Proj}_{\widehat{V}_k} E_k\| \\ &\stackrel{(c)}{\leq} \|\widetilde{\mathbb{B}}_k - \widetilde{B}_k\| + \|E_k\| + \|\text{Proj}_{\widehat{V}_k} E_k\| \\ &\leq 3\|E_k\|, \end{aligned} \tag{56}$$

where inequality (a) (b) are simply triangle inequality; and inequality (c) used the fact that  $\text{Proj}_{\widehat{V}_k} \widetilde{B}_k \text{Proj}_{\widehat{V}_k}$  from truncated SVD is the best rank  $R$  approximation to  $\widetilde{B}_k$  that minimizes the spectral norm. Finally, apply Lemma C.1 we have

$$\|\text{Proj}_{\widehat{V}_k} \widetilde{\mathbb{B}}_{\mathcal{I}_k}^{sqr t} - \widetilde{\mathbb{B}}_{\mathcal{I}_k}^{sqr t}\| \leq \sqrt{\|\text{Proj}_{\widehat{V}_k} \widetilde{\mathbb{B}}_k \text{Proj}_{\widehat{V}_k} - \widetilde{\mathbb{B}}_k\|}.$$

$\square$

**Lemma C.1.** Let  $U$  be a matrix of dimension  $M \times R$ . Let  $P$  be a projection matrix, we have

$$\|U - PU\|^2 \leq \|UU^\top - PU(PU)^\top\|.$$

*Proof.* (to Lemma C.1 ) Let  $P^\perp = I - P$ , so  $U - PU = P^\perp U$ . We can write

$$\begin{aligned} UU^\top - PU(PU)^\top &= (P + P^\perp)UU^\top(P + P^\perp) - PU(PU)^\top \\ &= P^\perp UU^\top P^\perp + PUU^\top P^\perp + P^\perp UU^\top P. \end{aligned}$$

Let vector  $v$  denote the leading left singular vector of  $P^\perp U$  and  $P^\perp U$ , by orthogonal projection it must be that  $Pv = 0$ . We can bound

$$\begin{aligned} \|UU^\top - PU(PU)^\top\| &\geq |v^\top (P^\perp UU^\top P^\perp + PUU^\top P^\perp + P^\perp UU^\top P)v| \\ &= |v^\top P^\perp UU^\top P^\perp v| \\ &= \|P^\perp U\|^2. \end{aligned}$$

□

**Lemma C.2** (Scaled noise matrix). *Consider a noise matrix  $E_S$  with independent entries, and each entry has sub-exponential tail with parameter  $(\sigma_{i,j} = \frac{1}{\sqrt{N}}, b_{i,j} = \frac{1}{N\sqrt{\rho_i\rho_j}})$ , for  $b_{i,j} \leq \frac{M}{N}$ .*

*Consider a fixed matrix  $V$  of dimension  $M \times R$  whose columns are orthonormal, with large probability we can bound the norm of  $V^\top E_S V$  and  $V^\top E_S$  separately by:*

$$\|V^\top E_S V\| = O(\sqrt{\frac{R^2}{M}}), \quad \text{and} \quad \|\hat{V}^\top E_S\| = O(\sqrt{\frac{RM}{N}}).$$

*Proof.* To bound the norm of the projected matrix, note that we have

$$\|V^\top E_S V\|_2^2 \leq \|V^\top E_S V\|_F^2 = \text{Tr}(V^\top E_S V V^\top \tilde{E}^\top V).$$

By Markov inequality, we have

$$\Pr(\text{Tr}(\hat{V}^\top E_S V V^\top E_S^\top \hat{V}) > t) \leq \frac{1}{t} \mathbb{E} \text{Tr}(\hat{V}^\top E_S V V^\top E_S^\top \hat{V}) = \frac{1}{t} \text{Tr}(\hat{V}^\top \underbrace{\mathbb{E}[E_S V V^\top E_S^\top]}_X \hat{V}) = \frac{1}{t} \frac{R^2}{N},$$

where the last equality is because for the  $i, j$ -th entry of  $X$  (let  $E_i$  denote the  $i$ -th row of  $E$  and  $V_r$  denote the  $r$ -th column of  $V$ )

$$X_{i,j} = \mathbb{E}[\sum_r (E_i V_r)(E_j V_r)] = \delta_{i,j} \sigma_{i,j}^2 \sum_r \|V_r\|_2^2 = \delta_{i,j} \frac{R}{N}.$$

Therefore, with probability at least  $1 - \delta$ , we have

$$\|V^\top E_S V\| \leq \sqrt{\frac{R^2}{N\delta}}.$$

Similarly, note that

$$\|\hat{V}^\top E_S\|_2^2 \leq \|\hat{V}^\top E_S\|_F^2 = \text{Tr}(\hat{V}^\top E_S E_S^\top \hat{V}).$$

By Markov inequality, we have

$$\Pr(\text{Tr}(\hat{V}^\top E_S E_S^\top \hat{V}) > t) \leq \frac{1}{t} \mathbb{E} \text{Tr}(\hat{V}^\top \tilde{E} E_S^\top \hat{V}) = \frac{1}{t} \text{Tr}(\hat{V}^\top \mathbb{E}[E_S E_S^\top] \hat{V}) = \frac{1}{t} \frac{RM}{N}.$$

□

*Proof.* (to Lemma 4.6 )

We first show the spectral concentration of  $\text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}}$  to  $D_S \tilde{\mathbb{B}} D_S$  can be bounded as:

$$\|\text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}} - D_S \tilde{\mathbb{B}} D_S\| = O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/M}\right)^{1/4}. \quad (57)$$

Note that by definition the rows and columns that are set to zero do not necessarily coincide in  $\tilde{B}$  and  $\tilde{\mathbb{B}}$  (defined in (37)), and we do not observe the sparsity pattern in  $\tilde{\mathbb{B}}$ .

Define  $\tilde{E} = \tilde{B} - \tilde{\mathbb{B}}$ . Apply triangle inequalities we have

$$\|\text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}} - D_S \tilde{\mathbb{B}} D_S\| \leq \|\text{Proj}_{\hat{V}} D_S \tilde{\mathbb{B}} D_S \text{Proj}_{\hat{V}} - D_S \tilde{\mathbb{B}} D_S\| + \|\text{Proj}_{\hat{V}} D_S \tilde{E} D_S \text{Proj}_{\hat{V}}\| \quad (58)$$

Next, we bound the two terms in (58) separately.

(1) First, to bound the term  $\|\text{Proj}_{\hat{V}} D_S \tilde{\mathbb{B}} D_S \text{Proj}_{\hat{V}} - D_S \tilde{\mathbb{B}} D_S\|$ , we note that

$$\|D_S \tilde{\mathbb{B}}^{sqr t}\| \leq \|\text{diag}(\rho^{-1/2}) \tilde{\mathbb{B}} \text{diag}(\rho^{-1/2})\| = 1. \quad (59)$$

Apply the block concentration result in Lemma 4.5, and recall that  $d_k^{max} = M_k \bar{\rho}_k^2 / w_{\min} = W_k \rho_k / w_{\min}$ , we have

$$\begin{aligned} \|\text{Proj}_{\hat{V}} D_S \tilde{\mathbb{B}}^{sqr t} - D_S \tilde{\mathbb{B}}^{sqr t}\| &\leq \left(\sum_k \bar{\rho}_k^{-1} \|\text{Proj}_{\hat{V}_k} \tilde{\mathbb{B}}_{\mathcal{I}_k}^{sqr t} - \tilde{\mathbb{B}}_{\mathcal{I}_k}^{sqr t}\|^2\right)^{1/2} \\ &= O\left(\left(\sum_k \frac{\sqrt{N d_k^{max} \log N d_k^{max}}}{N \bar{\rho}_k}\right)^{1/2}\right) \\ &\leq O\left(\left(\frac{1}{\sqrt{N}} \sum_k \sqrt{\log\left(\frac{N W_k \bar{\rho}_k}{w_{\min}}\right) \frac{M_k}{w_{\min}}}\right)^{1/2}\right) \\ &= O\left(\left(\sqrt{\frac{M}{N w_{\min}}} \sum_k \sqrt{\log\left(\frac{N W_k e^k}{M w_{\min}}\right) W_k e^{-k}}\right)^{1/2}\right) \\ &= O\left(\left(\sqrt{\frac{M}{N w_{\min}}} \sum_k \sqrt{(\log\left(\frac{N}{M w_{\min}}\right) + \log(W_k e^k)) W_k e^{-k}}\right)^{1/2}\right) \\ &\leq O\left(\left(\sqrt{\frac{M}{N w_{\min}}} \log \frac{N}{M w_{\min}} \sum_k \sqrt{\log(W_k e^k) W_k e^{-k}}\right)^{1/2}\right) \\ &= O\left(\left(\frac{\log(Nw_{\min}^2/M) + 3 \log(1/w_{\min})}{Nw_{\min}/M}\right)^{1/4}\right), \\ &= O\left(\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/M}\right)^{1/4}\right), \end{aligned} \quad (60)$$

where the last inequality is because  $\log(1/w_{\min}) \leq 1/w_{\min}$ , and the second last inequality is because

$$\sum_k \sqrt{\log(W_k e^k) W_k e^{-k}} \leq \sum_k \sqrt{W_k k e^{-k}} \leq \sqrt{\sum_k W_k \sum_k k e^{-k}} \leq 2.$$

Therefore, with (59) and (60) we can bound the first term in (58) by:

$$\begin{aligned}
& \|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}} D_S \text{Proj}_{\hat{\mathcal{V}}} - D_S \tilde{\mathbb{B}} D_S\| \\
& \leq \|(\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr} - D_S \tilde{\mathbb{B}}^{sqr})(\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr})^\top\| + \|D_S \tilde{\mathbb{B}}^{sqr}(\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr} - \tilde{\mathbb{B}}^{sqr})^\top\| \\
& \leq 2\|D_S \tilde{\mathbb{B}}^{sqr}\| \|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{\mathbb{B}}^{sqr} - D_S \tilde{\mathbb{B}}^{sqr}\| \\
& = O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/M}\right)^{1/4}.
\end{aligned} \tag{61}$$

(2) Second, to bound  $\|\text{Proj}_{\hat{\mathcal{V}}} D_S \tilde{E} D_S \text{Proj}_{\hat{\mathcal{V}}}\|$ , we carefully analyze the regularization to take care of the spillover effect. In Figure 4 we divide  $\tilde{E}$  into different regions according to the sparsity pattern of  $\tilde{\mathbb{B}}$  (as defined in (37)) and the regularized empirical matrix  $\tilde{B}$  in this step. We only highlight the division in one diagonal block, but it applies to the entire matrix across different bins. We bound the spectral norm of the matrix  $\tilde{E}$  restricting to different regions separately.

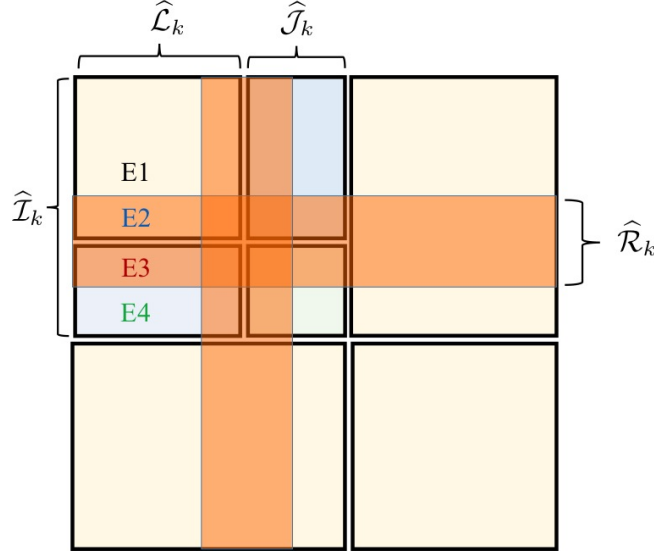


Figure 4: decomposition of  $\tilde{E}$  corresponding to  $\hat{\mathcal{I}}_k$ ,  $\hat{\mathcal{L}}_k$  and  $\hat{\mathcal{R}}_k$ .

In particular, region  $E_1$  is where rows/columns are not removed by either  $\tilde{\mathbb{B}}$  or  $\tilde{B}$ .

The entries are dominated by independent variables  $\frac{1}{N\sqrt{\rho_i\rho_j}}(\text{Poi}(N\mathbb{B}_{i,j}) - N\mathbb{B}_{i,j})$  and have sub-exponential tail with parameter  $(\sigma_{i,j} = \frac{1}{\sqrt{N}}, b_{i,j} = \frac{1}{N\sqrt{\rho_i\rho_j}} < 1)$ .

Also, since independent copies of the empirical bigram matrix are used in each step of the algorithm, the noise is independent with the  $(R \log M)$ -dimensional projection matrix  $\text{Proj}_{\hat{\mathcal{V}}}$ . Apply Lemma C.2, with probability at least  $1 - \delta$  we can bound the norm of projected noise as:

$$\|\text{Proj}_{\hat{\mathcal{V}}} D_S E_1 D_S \text{Proj}_{\hat{\mathcal{V}}}\| \leq 2\|\text{Proj}_{\hat{\mathcal{V}}} E_S \text{Proj}_{\hat{\mathcal{V}}}\| = O\left(\sqrt{\frac{(R \log M)^2}{N\delta}}\right) = o(1)$$

Region  $E_3$  corresponds to the rows/columns that are removed by both  $\tilde{\mathbb{B}}$  and  $\tilde{B}$ , thus  $E_3 = 0$ .

Region  $E_2$  is set to 0 in  $\tilde{B}$  but not in  $\tilde{\mathbb{B}}$ , thus the entries of  $E_2$  are equal to  $[\tilde{\mathbb{B}}]_{i,j}$ . For the rows of  $\tilde{\mathbb{B}}^{sqr}$  restricted to bin  $\hat{\mathcal{I}}_k$ , the row sums are bounded by  $2\bar{\rho}_k$ , and the column sums are bounded by  $W_k^s = O(W_k e^{-e^k d_0/2})$  (Lemma 4.3). Recall the fact that if a matrix  $X$  in which each row has  $\mathcal{L}_1$  norm



at most  $a$  and each column has  $\mathcal{L}_1$  norm at most  $b$ , then  $\|X\|_2 \leq \sqrt{ab}$ . Therefore, we can bound

$$\begin{aligned}\|\text{Proj}_{\hat{V}} D_S E_2 D_S \text{Proj}_{\hat{V}}\| &\leq \sum_k \left( \frac{1}{\sqrt{\bar{\rho}_k}} \sqrt{W_k^s \bar{\rho}_k} \right)^2 \\ &= \sum_k W_k^s \\ &= O(e^{-N/2M}).\end{aligned}\tag{62}$$

Region  $E_4$  is set to 0 in  $\tilde{\mathbb{B}}$  but not in  $\tilde{B}$ , corresponding to a subset of spillover words, and  $E_4 = \tilde{B}_4$ . There are at most  $\bar{W}_k/\bar{\rho}_k$  rows of region  $E_4$  in each bin  $\hat{\mathcal{I}}_k$ . Moreover, the row/column sum in are bounded by  $2\bar{\rho}_k$ . Conditional on the row sum, the entries in the row are distributed as multinomial  $\text{Mul}(\rho; 2\bar{\rho}_k)$ , thus the entries of  $D_S E_4 D_S$  are dominated by subexponential tail with parameter ( $\sigma_{i,j} = \frac{1}{\sqrt{N}}, b_{i,j} = \frac{1}{N\sqrt{\bar{\rho}_i \bar{\rho}_j}} < 1$ ). With probability at least  $\delta$  we can bound

$$\begin{aligned}\|\text{Proj}_{\hat{V}} D_S E_4 D_S \text{Proj}_{\hat{V}}\| &\leq \|\text{Proj}_{\hat{V}} [E_S]_4 \text{Proj}_{\hat{V}} + \text{Proj}_{\hat{V}} D_S (\bar{\rho}_k \mathbf{1} \rho^\top) D_S \text{Proj}_{\hat{V}}\| \\ &\leq \|\text{Proj}_{\hat{V}} [E_S]_4 \text{Proj}_{\hat{V}}\| + \|\text{Proj}_{\hat{V}} D_S (\bar{\rho}_k \mathbf{1} \rho^\top) D_S \text{Proj}_{\hat{V}}\| \\ &\leq \sqrt{\frac{\min(\sum_k \frac{\bar{W}_k}{\bar{\rho}_k}, R \log M)(R \log M)}{N\delta}} + \left( \sum_k \left( \frac{1}{\sqrt{\bar{\rho}_k}} \sqrt{W_k \bar{\rho}_k} \right)^2 \right)^{1/2} \\ &= O\left( \sqrt{\frac{(R \log M)^2}{N\delta}} + e^{-N/2M} \right),\end{aligned}$$

where the second last inequality is by the same argument as that in (62).

By triangle inequality over the 4 different regions, we can bound:

$$\|\text{Proj}_{\hat{V}} D_S \tilde{E} D_S \text{Proj}_{\hat{V}}\| = O\left( \sqrt{\frac{(R \log M)^2}{N\delta}} + e^{-N/2M} \right).\tag{63}$$

Therefore, with (61) and (63) we can bound (58) by:

$$\|\text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}} - D_S \tilde{\mathbb{B}} D_S\| = O\left( \frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/M} \right)^{1/4}.$$

Finally note that  $\hat{B}_1$  is the best rank  $R$  approximation of  $\text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}}$  and that  $D_S \tilde{\mathbb{B}} D_S$  is of rank at most  $R$ . We have

$$\begin{aligned}\|\hat{B}_1 - D_S \tilde{\mathbb{B}} D_S\| &\leq \|\hat{B}_1 - \text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}}\| + \|\text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}} - D_S \tilde{\mathbb{B}} D_S\| \\ &\leq 2\|\text{Proj}_{\hat{V}} D_S \tilde{B} D_S \text{Proj}_{\hat{V}} - D_S \tilde{\mathbb{B}} D_S\|.\end{aligned}$$

□

*Proof.* (to Lemma 4.7) By triangle inequality we have

$$\|\hat{B}_2 - \mathbb{B}\|_1 \leq \|\hat{B}_2 - \tilde{\mathbb{B}}\|_1 + \|\tilde{\mathbb{B}} - \mathbb{B}\|_1\tag{64}$$

Note that

$$\|\tilde{\mathbb{B}} - \mathbb{B}\|_1 \leq 2 \sum_k W_k^s = 2 \sum_k e^{-e^k N/2M} = o\left( \frac{MR^2}{Nw_{\min}^2} \right)^{1/2}.$$

Apply Cauchy-Schwartz to the first term we have:

$$\begin{aligned}
\|\widehat{B}_2 - \mathbb{B}\|_1 &= \sum_{i,j} |(\widehat{B}_2)_{i,j} - \mathbb{B}_{i,j}| \frac{1}{\sqrt{\rho_i \rho_j}} \sqrt{\rho_i \rho_j} \\
&\leq \|D_S(\widehat{B}_2 - \mathbb{B})D_S\|_F \sqrt{\sum_{i,j} \rho_i \rho_j} \\
&\leq \sqrt{2R} \|D_S(\widehat{B}_2 - \mathbb{B})D_S\| \\
&= O\left(\frac{\log(Nw_{\min}^2/M)}{Nw_{\min}^2/MR^2}\right)^{1/4}.
\end{aligned}$$

where the second inequality used the fact that if a matrix  $X$  if of rank  $R$  then  $\|X\|_F \leq \sqrt{R}\|X\|$ .  $\square$

*Proof.* (to Lemma 4.8 ) After removing the abnormally heavy rows and columns, we have that each row /column corresponding to a word in  $\widehat{\mathcal{I}}_k$  (defined according to Step 1 binning) has row sum and column sum less than  $2\bar{\rho}_k$ . We have that entries of  $E_S =: (D_S \widehat{B} D_S - D_S \mathbb{B} D_S)$  are dominated by entry-wise independent zero mean sub-exponential variable with parameter  $(\sigma_{i,j} = \frac{1}{\sqrt{N}}, b_{i,j} = \frac{1}{N\sqrt{\rho_i \rho_j}})$ , and  $b_{i,j} \leq \frac{M}{N} < 1$ . Given the initialization  $\widehat{B}_1$  from Step 3 such that  $\|\widehat{B}_1 - D_S \mathbb{B} D_S\| < \frac{1}{4}\sigma_{\min}(D_S \mathbb{B} D_S)$ , the correctness of Step 4 of Algorithm 2 follows Lemma C.3 below.  $\square$

**Lemma C.3** (Refinement with separation condition). *Consider a noisy low rank matrix  $X = UU^\top + E_S$ . Assume that the noise entries are zero mean, independent and  $\mathbb{E}[E_S]_{i,j}^2 \leq \frac{1}{N}$ . Assume that  $\sigma_{\min}(U) > (MR^2/N)^{1/8}$ . Given initialization  $\widehat{U}$  such that  $\|\widehat{U}\widehat{U}^\top - UU^\top\| = \epsilon_0 \leq \frac{1}{4}\sigma_{\min}(U)^2$ . We can find  $\widehat{X}$  such that*

$$\|\widehat{X} - X\|_F = O\left(\sqrt{\frac{MR}{N}}\right).$$

*Proof.* Let  $\widehat{V}$  and  $V$  denote the leading left singular vectors of  $\widehat{U}$  and  $U$ .

$$\|\text{Proj}_{\widehat{V}^\perp} UU^\top \text{Proj}_{\widehat{V}^\perp}\| = \|\text{Proj}_{\widehat{V}^\perp} (\widehat{U}\widehat{U}^\top - UU^\top) \text{Proj}_{\widehat{V}^\perp}\| \leq \epsilon_0.$$

First, consider the  $R \times R$  matrix  $\widehat{V}^\top X \widehat{V}$ , we know that with large probability,

$$\|\widehat{V}^\top X \widehat{V} - \widehat{V}^\top UU^\top \widehat{V}\| = \|\widehat{V}^\top E_S \widehat{V}\| = O\left(\sqrt{\frac{R^2}{N}}\right).$$

Let  $Z = (\widehat{V}^\top X \widehat{V})^{1/2}$ , we know that there exists some unknown rotation matrix  $H_Z$  such that

$$\|Z - \widehat{V}^\top U H_Z\| = o(1).$$

Note that  $U = \text{Proj}_{\widehat{V}} U + \text{Proj}_{\widehat{V}^\perp} U$ , we have  $\sigma_{\min}(U) \leq \sigma_{\min}(\widehat{V}^\top U) + \sigma_{\max}(\text{Proj}_{\widehat{V}^\perp} U)$ . By assumption of  $\epsilon_0$  we have

$$\sigma_{\min}(Z) = \sigma_{\min}(\widehat{V}^\top U) \geq \sigma_{\min}(U) - \epsilon_0^{1/2} \geq \frac{1}{2}\sigma_{\min}(U).$$

Next, consider the matrix  $\widehat{V}^\top X$  we know that it can be factorized as:

$$\begin{aligned}\widehat{V}^\top X &= \widehat{V}^\top (UU^\top + E_S) \\ &= \widehat{V}^\top UH_Z(UH_Z)^\top + \widehat{V}^\top E_S \\ &= Z(UH_Z)^\top + \widehat{V}^\top E_S + o(1).\end{aligned}$$

Let  $\widehat{U} = (Z^{-1}\widehat{V}^\top X)^\top$ . Note that  $\widehat{U} - UH_Z = (Z^{-1}\widehat{V}^\top E_S)^\top$ . Thus we can bound that

$$\begin{aligned}\|\widehat{U}\widehat{U}^\top - UU^\top\|_F &= \|\widehat{U}\widehat{U}^\top - UH_Z(UH_Z)^\top\|_F \\ &\leq \|(\widehat{U} - UH_Z)(UH_Z)^\top\|_F + \|(\widehat{U} - UH_Z)\widehat{U}^\top\|_F \\ &\leq \|UH_Z Z^{-1}\widehat{V}^\top E_S\|_F + \|\widehat{U} Z^{-1}\widehat{V}^\top E_S\|_F\end{aligned}$$

We bound the two terms separately. First

$$\begin{aligned}\|UH_Z Z^{-1}\widehat{V}^\top E_S\|_F &\leq \|\widehat{V}^\top UH_Z Z^{-1}\widehat{V}^\top E_S\|_F + \|(\widehat{V}^\perp)^\top UH_Z Z^{-1}\widehat{V}^\top E_S\|_F \\ &\leq \|ZZ^{-1}\widehat{V}^\top E_S\|_F + \|(\widehat{V}^\perp)^\top U\| \|Z^{-1}\| \|\widehat{V}^\top E_S\|_F \\ &\leq \|\widehat{V}^\top E_S\|_F (1 + \sqrt{\epsilon_0}/\sigma_{\min}) \\ &\leq 2\|\widehat{V}^\top E_S\|_F\end{aligned}$$

We then bound the second term

$$\begin{aligned}\|\widehat{U} Z^{-1}\widehat{V}^\top E_S\|_F &\leq \|UH_Z Z^{-1}\widehat{V}^\top E_S\|_F + \|(Z^{-1}V^\top E_S)^\top Z^{-1}\widehat{V}^\top E_S\|_F \\ &\leq \|\widehat{V}^\top E_S\|_F (2 + \|\widehat{V}^\top E_S\|_F / \sigma_{\min}(Z)^2) \\ &\leq 6\|\widehat{V}^\top E_S\|_F,\end{aligned}$$

where the last inequality is by the assumption  $\sigma_{\min}(U)^2 > (MR/N)^{1/4} > (MR/N)^{1/2} = \|\widehat{V}^\top E_S\|_F$ .

Finally by Lemma C.2 we have that with large probability

$$\|\widehat{U}\widehat{U}^\top - UU^\top\|_F \leq 8\|\widehat{V}^\top E_S\|_F = O(\sqrt{\frac{MR}{N}}).$$

□

## D Proofs for HMM testing lower bound

*Proof.* (to Theorem 5.1)

$$\begin{aligned}
TV(\Pr_1(G_1^N), \Pr_2(G_1^N)) &\leq TV(\Pr_1(G_1^N, \mathcal{A}), \Pr_2(G_1^N, \mathcal{A})) \\
&= \frac{1}{2} \sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{M}{M/2}} |\Pr_2(G_1^N, \mathcal{A}) - \Pr_1(G_1^N, \mathcal{A})| \\
&= \frac{1}{2} \sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{M}{M/2}} \Pr_1(G_1^N, \mathcal{A}) \left| \frac{\Pr_2(G_1^N, \mathcal{A})}{\Pr_1(G_1^N, \mathcal{A})} - 1 \right| \\
&\stackrel{(a)}{\leq} \frac{1}{2} \left( \sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{M}{M/2}} \Pr_1(G_1^N, \mathcal{A}) \left( \frac{\Pr_2(G_1^N, \mathcal{A})}{\Pr_1(G_1^N, \mathcal{A})} - 1 \right)^2 \right)^{1/2} \\
&= \frac{1}{2} \left( \sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{M}{M/2}} \Pr_1(G_1^N, \mathcal{A}) \left( \frac{\Pr_2(G_1^N, \mathcal{A})}{\Pr_1(G_1^N, \mathcal{A})} \right)^2 - 1 \right)^{1/2} \\
&\stackrel{(b)}{=} \frac{1}{2} \left( \left( \frac{M^N}{\binom{M}{M/2}} \right)^2 \sum_{G_1^N \in [M]^N, \mathcal{A} \in \binom{M}{M/2}} \Pr_1(G_1^N, \mathcal{A}) \left( \sum_{\mathcal{B} \in \binom{M}{M/2}} \Pr_2(G_1^N | \mathcal{B}) \right)^2 - 1 \right)^{1/2} \\
&\stackrel{(c)}{=} \frac{1}{2} \left( \underbrace{\frac{M^N}{\binom{M}{M/2}^2} \sum_{G_1^N \in [M]^N} \left( \sum_{\mathcal{B} \in \binom{M}{M/2}} \Pr_2(G_1^N | \mathcal{B}) \right)^2}_{Y} - 1 \right)^{1/2}, \tag{66}
\end{aligned}$$

where inequality (a) used the fact that  $\mathbb{E}[X] \leq (\mathbb{E}[X^2])^{1/2}$ ; equality (b) used the joint distributions in (46) and (49); and equality (c) takes sum over the summand  $\mathcal{A}$  first and makes use of the marginal probability  $\Pr_1(G_1^N)$  as in (48).

In order to bound the term  $Y = \frac{M^N}{\binom{M}{M/2}^2} \sum_{G_1^N \in [M]^N} \left( \sum_{\mathcal{B} \in \binom{M}{M/2}} \Pr_2(G_1^N | \mathcal{B}) \right)^2$  in equation (66), we break the square and write:

$$Y = \frac{M^N}{\binom{M}{M/2}^2} \sum_{\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}} \sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}'). \tag{67}$$

In Claim D.1 below we explicitly compute the term  $\sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}')$  for any two subsets  $\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}$ . Then in Claim D.2 we compute the sum over  $\mathcal{B}, \mathcal{B}'$  and bound  $Y \leq \frac{1}{\sqrt{1 - \frac{2(1-2t)N}{M}}}$ .

To conclude, we can bound the total variation distance as:

$$TV(\Pr_1(G_1^N), \Pr_2(G_1^N)) \leq \frac{1}{2} \sqrt{\frac{1}{\sqrt{1 - \frac{2(1-2t)N}{M}}} - 1}.$$

In the case that  $N \leq M$ , this is bounded as  $TV(\Pr_1(G_1^N), \Pr_2(G_1^N)) \leq \frac{\sqrt{1-2t}}{2} \sqrt{\frac{N}{M}}$ , which vanishes as  $N = o(M)$  for any constant transition probability  $t$ .

□

**Claim D.1.** *In the same setup of Theorem 5.1, given two subsets  $\mathcal{B}, \mathcal{B}' \in \mathcal{M}$  and  $|\mathcal{B}| = |\mathcal{B}'| = M/2$ , let  $\bar{\mathcal{B}} = \mathcal{M} \setminus \mathcal{B}, \bar{\mathcal{B}}' = \mathcal{M} \setminus \mathcal{B}'$  denote the corresponding complement. Define  $x \in [0, 1]$  to be:*

$$x = |\mathcal{B} \cap \mathcal{B}'| / (M/2). \quad (68)$$

Let  $\gamma_1(x) = \frac{1}{2} \left( 1 + (1 - 2t)^2 + \sqrt{(1 - (1 - 2t)^2)^2 + (2(1 - 2t))^2 x^2} \right)$  and let  $\gamma_2(x) = \frac{1}{2} \left( 1 + (1 - 2t)^2 - \sqrt{(1 - (1 - 2t)^2)^2 + (2(1 - 2t))^2 x^2} \right)$  be functions of  $|\mathcal{B} \cap \mathcal{B}'|$  and  $t$ . We have:

$$\sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}') = \frac{1}{(M/2)^N} \left( \frac{\gamma_1(x)^N (1 - 2\gamma_2(x)) - \gamma_2(x)^N (1 - 2\gamma_1(x))}{2(\gamma_1(x) - \gamma_2(x))} \right),$$

*Proof.* (to Claim D.1)

(1) For an instance of 2-state HMM which support for  $q$  is specified by set  $\mathcal{B} \in \binom{M}{M/2}$ , consider two consecutive outputs  $(g_{n-1}, g_n)$ . We first show how to compute the probability  $\Pr_2(g_n | g_{n-1}, \mathcal{B})$ .

Given  $\mathcal{B}$  and another set  $\mathcal{B}'$ , we can partition the vocabulary  $\mathcal{M}$  into four subsets as:

$$\mathcal{M}_1 = \mathcal{B} \cap \mathcal{B}', \quad \mathcal{M}_2 = \mathcal{B} \cap \bar{\mathcal{B}}', \quad \mathcal{M}_3 = \bar{\mathcal{B}} \cap \mathcal{B}', \quad \mathcal{M}_4 = \bar{\mathcal{B}} \cap \bar{\mathcal{B}}'.$$

Note that we have  $|\mathcal{M}_1| = |\mathcal{M}_4| = xM/2$  and  $|\mathcal{M}_2| = |\mathcal{M}_3| = (1 - x)M/2$ .

Define a subset of tuples  $\mathcal{J}_B \subset [4]^2$  to be

$$\mathcal{J}_B = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 3), (3, 4), (4, 3), (4, 4)\}, \quad \mathcal{J}_B^c = [4]^2 \setminus \mathcal{J}_B.$$

If  $g_{n-1} \in \mathcal{M}_{j'}, g_n \in \mathcal{M}_j$  and  $(j', j) \in \mathcal{J}_B$ , we know that the hidden state for the HMM associated with set  $\mathcal{B}$  does not change between time slot  $n - 1$  and  $n$ , namely  $s_{n-1} = s_n$ . Thus  $\Pr_2(g_n | g_{n-1}, \mathcal{B}) = \Pr_2(s_n | s_{n-1}, \mathcal{B}) \Pr_2(g_n | s_n, \mathcal{B}) = \frac{1-t}{M/2}$ . Also, if  $(j', j) \in \mathcal{J}_B^c$ , we know that there is the state transition and we have  $\Pr_2(g_n | g_{n-1}, \mathcal{B}) = \frac{t}{M/2}$ .

Similarly, for the 2-state HMM associated with set  $\mathcal{B}'$ , we can define the set of tuples

$$\mathcal{J}_{B'} = \{(1, 1), (1, 3), (3, 1), (3, 3), (2, 2), (2, 4), (4, 2), (4, 4)\}, \quad \mathcal{J}_{B'}^c = [4]^2 \setminus \mathcal{J}_{B'}.$$

Here  $\Pr_2(g_n | g_{n-1}, \mathcal{B}') = \frac{1-t}{M/2}$  if  $(j', j) \in \mathcal{J}_{B'}$  and equals  $\frac{t}{M/2}$  if  $(j', j) \in \mathcal{J}_{B'}^c$ .

(2) Next, we show how to compute the target sum of the claim statement in a recursive way.

For fixed sets  $\mathcal{B}$  and  $\mathcal{B}'$ , define  $F_{n,j}$  for  $n \leq N$  and  $j = 1, 2, 3, 4$  as below

$$F_{n,j} = \sum_{G_1^n \in [M]^n} \Pr_2(G_1^n | \mathcal{B}) \Pr_2(G_1^n | \mathcal{B}') \mathbf{1}[g_n \in \mathcal{M}_j],$$

and the target sum is  $\sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}') = \sum_{j=1:4} F_{N,j}$ . Also, we have that

$$F_{1,j} = |\mathcal{M}_j| / M^2.$$

Making use of the recursive property of the probability rule of the 2-state HMM as in (47), we can

write the following recursion in terms of  $F_{n,j}$  for  $n \geq 2$ :

$$\begin{aligned}
& F_{n,j} \\
&= \sum_{G_1^n \in [M]^n} \Pr_2(G_1^n | \mathcal{B}) \Pr_2(G_1^n | \mathcal{B}') \mathbf{1}[g_n \in \mathcal{M}_j] \\
&= \sum_{G_1^n \in [M]^n} \Pr_2(G_1^{n-1} | \mathcal{B}) \Pr_2(G_1^{n-1} | \mathcal{B}') \Pr_2(g_n | G_1^{n-1}, \mathcal{B}) \Pr_2(g_n | G_1^{n-1}, \mathcal{B}') \sum_{j'=1:4} \mathbf{1}[g_{n-1} \in \mathcal{M}_{j'}, g_n \in \mathcal{M}_j] \\
&= \sum_{G_1^{n-1} \in [M]^{n-1}} \Pr_2(G_1^{n-1} | \mathcal{B}) \Pr_2(G_1^{n-1} | \mathcal{B}') \sum_{g_n \in [M]} \sum_{j'=1:4} \mathbf{1}[g_{n-1} \in \mathcal{M}_{j'}, g_n \in \mathcal{M}_j] \Pr_2(g_n | g_{n-1}, \mathcal{B}) \Pr_2(g_n | g_{n-1}, \mathcal{B}') \\
&= |\mathcal{M}_j| \sum_{j'=1:4} F_{n-1,j'} \left( \frac{1-t}{M/2} \mathbf{1}[(j', j) \in \mathcal{J}_B] + \frac{t}{M/2} \mathbf{1}[(j, j') \in \mathcal{J}_B^c] \right) \left( \frac{1-t}{M/2} \mathbf{1}[(j', j) \in \mathcal{J}_{B'}] + \frac{t}{M/2} \mathbf{1}[(j, j') \in \mathcal{J}_{B'}^c] \right)
\end{aligned}$$

where we used the probability  $\Pr_2(g_n | g_{n-1}, \mathcal{B})$  derived in (1).

Equivalently we can write the recursion as:

$$\begin{pmatrix} F_{n,1} \\ F_{n,2} \\ F_{n,3} \\ F_{n,4} \end{pmatrix} = \frac{1}{(M/2)} D_x T \begin{pmatrix} F_{n-1,1} \\ F_{n-1,2} \\ F_{n-1,3} \\ F_{n-1,4} \end{pmatrix},$$

for diagonal matrix  $D_x = \begin{pmatrix} x & & & \\ & 1-x & & \\ & & 1-x & \\ & & & x \end{pmatrix}$  and the symmetric stochastic matrix  $T$  given by

$$T = \begin{pmatrix} (1-t)^2 & (1-t)t & (1-t)t & t^2 \\ (1-t)t & (1-t)^2 & t^2 & (1-t)t \\ (1-t)t & t^2 & (1-t)^2 & (1-t)t \\ t^2 & (1-t)t & (1-t)t & (1-t)^2 \end{pmatrix} = \sum_{i=1}^4 \lambda_i v_i v_i^\top,$$

where the singular values and singular vectors of  $T$  are specified as follows:  $\lambda_1 = 1$ ,  $\lambda_4 = (1-2t)^2$ , and  $v_1 = \frac{1}{2}[1, 1, 1, 1]^\top$ ,  $v_4 = \frac{1}{2}[1, -1, -1, 1]^\top$ . And  $\lambda_2 = \lambda_3 = 1-2t$  with  $v_2 = \frac{1}{\sqrt{2}}[0, 1, -1, 0]^\top$  and  $v_3 = \frac{1}{\sqrt{2}}[1, 0, 0, -1]^\top$ .

Note that we can write  $(F_{1,1}, F_{1,2}, F_{1,3}, F_{1,4})^\top = \frac{M/2}{M^2} D_x (1, 1, 1, 1)^\top$ .

(3) Finally we can compute the target sum as:

$$\begin{aligned}
\sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}') &= \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} F_{N,1} & F_{N,2} & F_{N,3} & F_{N,4} \end{pmatrix}^\top \\
&= \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \frac{1}{(M/2)^{N-1}} (D_x T)^{N-1} \frac{M/2}{M^2} D_x \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}^\top \\
&\stackrel{(a)}{=} \frac{1}{M^N} v_1^\top (2D_x T)^N v_1 \\
&= \frac{1}{M^N} \begin{pmatrix} 1 & 0 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & (2x-1) \\ (1-2t)^2(2x-1) & (1-2t)^2 \end{pmatrix}^N}_{H(x)^N} \begin{pmatrix} 1 & 0 \end{pmatrix}^\top \\
&\stackrel{(b)}{=} \frac{1}{M^N} \left( \frac{\gamma_1(x)\gamma_2(x)^N - \gamma_2(x)\gamma_1(x)^N}{\gamma_1(x) - \gamma_2(x)} + \frac{\gamma_1(x)^N - \gamma_2(x)^N}{\gamma_1(x) - \gamma_2(x)} \right) \\
&= \frac{1}{M^N} \frac{\gamma_1^N(1 - \gamma_2) - \gamma_2^N(1 - \gamma_1)}{\gamma_1 - \gamma_2},
\end{aligned}$$

where in (a) we used the fact that

$$2D_x T v_1 = v_1 + (2x-1)v_4, \text{ and } 2D_x T v_4 = (1-2t)^2((2x-1)v_1 + v_4).$$

In (b) we used the Calley-Hamilton theorem to obtain that for  $2 \times 2$  matrix  $H(x)$  parameterized by  $x$  and with 2 distinct eigenvalue  $\gamma_1(x)$  and  $\gamma_2(x)$ , its power can be written as  $H(x)^N = \frac{\gamma_1(x)\gamma_2(x)^N - \gamma_2(x)\gamma_1(x)^N}{\gamma_1(x) - \gamma_2(x)} I_{2 \times 2} + \frac{\gamma_1(x)^N - \gamma_2(x)^N}{\gamma_1(x) - \gamma_2(x)} H(x)$ . Moreover, the distinct eigenvalues of the  $2 \times 2$  matrix  $H(x)$  can be written explicitly as follows:

$$\gamma_1(x) = \frac{1}{2} \left( 1 + (1-2t)^2 + \sqrt{(1 - (1-2t)^2)^2 + (2(1-2t))^2 x^2} \right), \quad (69)$$

$$\gamma_2(x) = \frac{1}{2} \left( 1 + (1-2t)^2 - \sqrt{(1 - (1-2t)^2)^2 + (2(1-2t))^2 x^2} \right). \quad (70)$$

where recall that we defined  $x = \frac{|\mathcal{B} \cap \mathcal{B}'|}{M/2}$  so  $0 \leq x \leq 1$ , also we have the transition probability  $0 < t < 1/2$  to be a constant, therefore we have  $\gamma_1 > \gamma_2$  to be two distinct real roots.  $\square$

The next claim makes use of the above claim and bounds the right hand side of (67).

**Claim D.2.** *In the same setup of Theorem 5.1, we have*

$$Y = \frac{M^N}{\binom{M}{M/2}^2} \sum_{\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}} \sum_{G_1^N \in [M]^N} \Pr_2(G_1^N | \mathcal{B}) \Pr_2(G_1^N | \mathcal{B}') \leq \frac{1}{\sqrt{1 - \frac{2(1-2t)N}{M}}}.$$

*Proof.* (to Claim D.2)

Define  $f(x) = \frac{\gamma_1(x)^N(1-\gamma_2(x)) - \gamma_2(x)^N(1-\gamma_1(x))}{\gamma_1(x) - \gamma_2(x)}$  with  $\gamma_1(x)$  and  $\gamma_2(x)$  defined in (69) and (70) as functions of  $x$ . Recall that  $x = |\mathcal{B} \cap \mathcal{B}'|/(xM/2) \in [0, 1]$ .

Use the result of Claim D.1 we have:

$$\begin{aligned}
Y &= \frac{M^N}{\binom{M}{M/2}^2} \sum_{\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}} \frac{1}{M^N} f(x) \\
&\stackrel{(a)}{=} \frac{1}{\binom{M}{M/2}^2} \binom{M}{M/2} \sum_{i=1}^{M/2} \binom{M/2}{i}^2 f\left(\frac{i}{M/2}\right) \\
&= \frac{1}{\binom{M}{M/2}} \sum_{i=1}^{M/2} \binom{M/2}{i}^2 f\left(\frac{i}{M/2}\right), \tag{71}
\end{aligned}$$

where equality (a) is obtained by counting the number of subsets  $\mathcal{B}, \mathcal{B}' \in \binom{M}{M/2}$ : for each fixed  $\mathcal{B}$ , there are  $\binom{M/2}{i}$  choices of  $\mathcal{B}'$  such that  $|\mathcal{B} \cap \mathcal{B}'| = i$ .

Next we approximately bound  $Y$  when  $M$  is asymptotically large. First, note that  $\gamma_1(0) = 1$  and  $\gamma_1(1) = 1 + (1 - 2t)^2$ , we can bound  $\gamma_1(x)$  as by exponential function:

$$\gamma_1(x) = \frac{1}{2} \left( 1 + (1 - 2t)^2 + \sqrt{(1 - (1 - 2t)^2)^2 + (2(1 - 2t))^2 x^2} \right) \leq e^{(1-2t)(1-2x)^2},$$

Then note that for  $N$  increasing with  $M$  and thus asymptotically large, we have  $\gamma_2^N(1 - \gamma_1) = o(1)$ , so we bound  $f(x)$  by:

$$\lim_{M \rightarrow \infty} f(x) \approx \gamma_1(x)^N \frac{1 - \gamma_2(x)}{\gamma_1(x) - \gamma_2(x)} \leq e^{(1-2t)(1-2x)^2 N},$$

where we used the fact that  $1/2 \leq \gamma_1 \leq 1$  and that  $\frac{1 - \gamma_2(x)}{\gamma_1(x) - \gamma_2(x)} = \frac{1}{2} (1 + 1/\sqrt{1 + x^2 (\frac{2(1-2t)}{1-(1-2t)^2})^2}) \leq 1$ .

Second, we use Stirling's approximation for the combinatorial coefficients  $\binom{M/2}{i}^2$  and  $\binom{M}{M/2}$ ,

$$\begin{aligned}
\binom{M}{M/2} &\approx \frac{4^{M/2}}{\sqrt{\pi M/2}}, \\
\binom{M/2}{i}^2 &\approx \left( \binom{M/2}{M/4} e^{-(M/2 - 2i)^2 / 2(M/2)} \right)^2 \\
&\approx \frac{4^{M/2}}{\pi M/4} e^{-2M(i/(M/2) - 1/2)^2}, \quad \text{for } \log M \leq i \leq (M/2) - \log M.
\end{aligned}$$

Finally we can approximately bound  $Y$  in (71) as follows:

$$\begin{aligned}
Y &\approx \frac{1}{\sqrt{\pi M}} \frac{4}{\sqrt{2}} \sum_{i=1}^{M/2} e^{-2M(\frac{i}{M/2} - \frac{1}{2})^2} f\left(\frac{i}{M/2}\right) + \frac{2}{\binom{M}{M/2}} \sum_{i=1}^{\log M} \binom{M/2}{i} f\left(\frac{i}{M/2}\right) \\
&\leq \frac{1}{\sqrt{\pi M}} \frac{4}{\sqrt{2}} \left( \frac{M}{2} \int_{y=-1/2}^{1/2} e^{-2My^2 + 4(1-2t)y^2 N} dy \right) + o(1) \\
&= \sqrt{\frac{2M}{\pi}} \int_{y=-1/2}^{1/2} e^{-y^2 2M(1-2(1-2t)N/M)} dy \\
&= \frac{1}{\sqrt{1 - \frac{2(1-2t)N}{M}}},
\end{aligned}$$



where the second inequality is because for  $M$  asymptotically large and  $N = O(M)$ , we have

$$\frac{2}{\binom{M}{M/2}} \sum_{i=1}^{\log M} \binom{M/2}{i} f\left(\frac{i}{M/2}\right) \leq 2\sqrt{\pi(M/2)} 4^{-M/2} (\log M) (M/2)^{\log M} e^{4(1-2t)(2\log M/M)^2 N} = o(1).$$

□

## E Analyze truncated SVD

The reason that truncated SVD does not concentrate at the optimal rate is as follows. What truncated SVD actually optimizes is the spectral distance from the estimator to the empirical average (minimizing  $\|\hat{B} - \frac{1}{N}B_N\|_2$ ), yet not to the expected matrix  $\mathbb{B}$ . It is only “optimal” in some very special setup, for example when  $(\frac{1}{N}B_N - \mathbb{B})$  are entry-wise i.i.d. Gaussian. In the asymptotic regime when  $N \rightarrow \infty$  it is indeed true that under mild condition any sampling noise converges to i.i.d Gaussian. However in the sparse regime where  $N = \Omega(M)$ , the sampling noise from the probability matrix is very different from additive Gaussian noise.

**Claim E.1** (Truncated SVD has sample complexity super linear). *In order to achieve  $\epsilon$  accuracy, the sample complexity of rank-2 truncated SVD estimator is in given by  $N = O(M^2 \log M)$ .*

*Example 1:  $a = b = w = 1/2$ , dictionary given by*

$$p = \left[ \frac{1+C_\Delta}{M}, \dots, \frac{1+C_\Delta}{M}, \frac{1-C_\Delta}{M}, \dots, \frac{1-C_\Delta}{M} \right],$$

$$q = \left[ \frac{1-C_\Delta}{M}, \dots, \frac{1-C_\Delta}{M}, \frac{1+C_\Delta}{M}, \dots, \frac{1+C_\Delta}{M} \right].$$

*Sample complexity is  $O(M \log M)$ .*

*Example 2: modify Example 1 so that a constant fraction of the probability mass lies in a common word, namely  $p_1 = q_1 = 1/2\rho_1 = 0.1$ , while the marginal probability as well as the separation in all the other words are roughly uniform. Sample complexity is  $O(M^2 \log N)$ .*

*Proof.* (to Claim E.1 (Truncated SVD has sample complexity super linear))

(1) We formalize this and examine the sample complexity of t-SVD by applying Bernstein matrix inequality. The concentration of the empirical average matrix at the following rate:

$$\Pr(\|\frac{1}{N}B_N - \mathbb{B}\| \geq t) \leq e^{-\frac{(Nt)^2}{NV\text{ar} + BNt/3} + \log(M)},$$

where  $\text{Var} = \|\mathbb{E}[e_i e_i^\top]\|_2 = \|\text{diag}(\rho)\|_2 = \max_i \rho_i$ , and  $B = \max_{i,j} \|e_i e_j\|_2 = 1$ . Therefore, with probability at least  $1 - \delta$ , we have that

$$\|\frac{1}{N}B_N - \mathbb{B}\| \leq \sqrt{\frac{\max_i \rho_i \log(M/\delta)}{N}} + \frac{1}{3} \frac{1}{N} \log(M/\delta). \quad (72)$$

Since  $\|x\|_1 \leq \sqrt{M}\|x\|_2$ , in order to guarantee that  $\|\hat{\Delta} - \Delta\|_1 \leq \epsilon$ , it suffices to ensure that  $\|\hat{\Delta} - \Delta\|_2 \leq \epsilon/\sqrt{M}$ . Note that the leading two eigenvectors are given by  $\sigma_1(\mathbb{B}) \geq \|\rho\|_2 = 1/\sqrt{M}$  and  $\sigma_2(\mathbb{B}) = \|\Delta\|_2 = C_\Delta/\sqrt{M}$ . Assume that we have the exact marginal probability  $\rho$ , by Davis-Kahan, it suffices to ensure that

$$\|\frac{1}{N}B_N - \mathbb{B}\|_2 \leq \epsilon \frac{\|\Delta\|_2}{\sqrt{M}}.$$

Example 1. Consider the example of  $(p, q)$  in community detection problem, where the marginal probability  $\rho_i$  is roughly uniform. We have  $\|\Delta\|_2 = C_\Delta/\sqrt{M}$  and  $\max_i \rho_i = 1/M$ , and the concentration bound becomes

$$\left\| \frac{1}{N} B_N - \mathbb{B} \right\| \leq \sqrt{\frac{\log(M/\delta)}{MN}}, \quad (73)$$

and by requiring

$$\sqrt{\frac{\log(M/\delta)}{MN}} \leq \epsilon \frac{\|\Delta\|_2}{\sqrt{M}} = \epsilon \frac{C_\Delta}{M}$$

we get a sample complexity bound  $N = \Omega(M \log(M/\delta))$ , which is worse than the lower bound by a  $\log(M)$  factor.

Example 2. Moreover, modify Example 1 so that a constant fraction of the probability mass lies in a common word, namely  $p_1 = q_1 = 1/2\rho_1 = 0.1$ , while the marginal probability as well as the separation in all the other words are roughly uniform. In this case,  $\|\Delta\|_2$  is still roughly  $C_\Delta/\sqrt{M}$ , however we have  $\max_i \rho_i = 0.1$ , and the sample complexity becomes  $N = \Omega(M^2 \log(M/\delta))$ . This is even worse than the first example, as the same separation gets swamped by the heavy common words.

**(2)** (square root of the empirical marginal scaling (from 1st batch of samples) on both side of the empirical count matrix (from 2nd batch of samples)).  $\square$

Take a closer look at the above proof and we can identify two misfortunes that make the truncated SVD deviate from linear sample complexity:

1. In the worst case, the nonuniform marginal probabilities costs us an  $M$  factor in the first component of Bernstein's inequality;
2. We pay another  $\log(M)$  factor for the spectral concentration of the  $M \times M$  random matrix.

To resolve these two issues, the two corresponding key ideas of Phase I algorithm are “binning” and “regularization”:

1. “Binning” means that we partition the vocabulary according to the marginal probabilities, so that for the words in each bin, their marginal probabilities are roughly uniform. If we are able to apply spectral method in each bin separately, we could possibly get rid of the  $M$  factor.
2. Now restrict our attention to the diagonal block of the empirical average matrix  $\frac{1}{N} B_N$  whose indices corresponding to the words in a bin. Assume that the bin has sufficiently many words, so that the expected row sum and column sum are at least constant, namely the effective number of samples is at least in the order of the number of words in the bin.

We apply regularized spectral method for the empirical average with indices restricted to the bin. By “regularization” we mean removing the rows and column, whose row and column sum are much higher than the expected row sum, from the empirical. Then we apply t-SVD to the remaining. This regularization idea is motivated by the community detection literature in the sparse regime, where the total number of edges of the random network is only linear in the number of nodes.

## F Auxiliary Lemmas

**Lemma F.1** (Wedin's theorem applied to rank-1 matrix). *Denote symmetric matrix  $X = vv^\top + E$ . Let  $\widehat{v}\widehat{v}^\top$  denote the rank-1 truncated SVD of  $X$ . There is a positive universal constant  $C$  such that*

$$\min\{\|v - \widehat{v}\|, \|v + \widehat{v}\|\} \leq \begin{cases} \frac{C\|E\|}{\|v\|} & \text{if } \|v\|^2 > C\|E\|; \\ C\|E\|^{1/2} & \text{if } \|v\|^2 < C\|E\|. \end{cases}$$

**Lemma F.2** (Chernoff Bound for Poisson variables).

$$\begin{aligned} \Pr(\text{Poi}(\lambda) \geq x) &\leq e^{-\lambda} \left(\frac{x}{e\lambda}\right)^{-x}, \quad \text{for } x > \lambda, \\ \Pr(\text{Poi}(\lambda) \leq x) &\leq e^{-\lambda} \left(\frac{x}{e\lambda}\right)^{-x}, \quad \text{for } x < \lambda. \end{aligned}$$

**Lemma F.3** (Matrix Bernstein). *Consider a sequence of  $N$  random matrix  $\{X_k\}$  of dimension  $M \times M$  which are independent, self-adjoint. Assume that  $\mathbb{E}[X_k] = 0$  and  $\lambda_{\max}(X_k) \leq R$  almost surely. Denote the total variance by  $\sigma^2 = \|\sum_{k=1}^N \mathbb{E}[X_k^2]\|$ . Then the following inequality holds for all  $t > 0$ :*

$$\Pr\left(\left\|\sum_{k=1}^N X_k\right\| \geq t\right) \leq Me^{-\frac{t^2}{\sigma^2 + Rt/3}} \leq \begin{cases} Me^{-\frac{3t^2}{8\sigma^2}}, & \text{for } t \leq \sigma^2/R; \\ Me^{-\frac{3t}{8R}}, & \text{for } t \geq \sigma^2/R. \end{cases}$$

**Lemma F.4** (Upper bound of Poisson tails (Proposition 1 in [24])). *Assume  $\lambda > 0$ , consider the Poisson distribution  $\text{Poi}(\lambda)$ .*

(1) *if  $0 \leq n < \lambda$ , the left tail can be upper bounded by:*

$$\Pr(\text{Poi}(\lambda) \leq n) \leq \left(1 - \frac{n}{\lambda}\right)^{-1} \Pr(\text{Poi}(\lambda) = n).$$

(2) *if  $n > \lambda - 1$ , for any  $m \geq 1$ , the right tail can be upper bounded by:*

$$\Pr(\text{Poi}(\lambda) \geq n) \leq \left(1 - \left(\frac{\lambda}{n+1}\right)^m\right)^{-1} \sum_{i=n}^{n+m-1} \Pr(\text{Poi}(\lambda) = i).$$

**Corollary F.5.** *Let  $\lambda > C$  for some large universal constant  $C$ . For any constant  $c' > e$ ,  $0 \leq c < 1/2$ , we have the following Poisson tail bounds:*

$$\begin{aligned} \Pr(\text{Poi}(\lambda) \leq c\lambda) &\leq 2e^{-\lambda/2}, \\ \Pr(\text{Poi}(\lambda) \geq c'\lambda) &\leq 2e^{-c'\lambda}. \end{aligned}$$

*Proof.* Apply Stirling's bound for  $\lambda$  large, we have  $\lambda! \geq (\frac{\lambda}{e})^\lambda$ . Then, the bound in Lemma F.4 (1) can be written as

$$\begin{aligned} \Pr(\text{Poi}(\lambda) \leq c\lambda) &\leq (1 - c)^{-1} \Pr(\text{Poi}(\lambda) = c\lambda) \\ &\leq 2e^{-\lambda} (\lambda)^{c\lambda} / (c\lambda)! \\ &\leq 2e^{-\lambda} (\lambda)^{c\lambda} / (c\lambda e^{-1})^{c\lambda} \\ &\leq 2e^{-\lambda + c\lambda \log(e/c)} \\ &\leq 2e^{-\lambda/2}, \end{aligned}$$

where in the second inequality we used the assumption that  $c < 1/2$ , and in the last inequality we used the inequality  $1 - c \log(e/c) \geq 1/2$  for all  $0 \leq c < 1$ .

Similarly, set  $m = 1$  in Lemma F.4 (2), we can write the bound as

$$\begin{aligned}
\Pr(\text{Poi}(\lambda) \geq c'\lambda) &\leq (1 - \frac{\lambda}{c'\lambda + 1})^{-1} \Pr(\text{Poi}(\lambda) = c'\lambda) \\
&\leq (1 - 1/c')^{-1} e^{-\lambda} (\lambda)^{c'\lambda} / (c'\lambda)! \\
&\leq 2e^{-\lambda} (\lambda)^{c'\lambda} / (c'\lambda e^{-1})^{c'\lambda} \\
&\leq 2e^{-c'\lambda \log(c'/e) - 1} \\
&\leq 2e^{-c'\lambda},
\end{aligned}$$

where in both the second and the last inequality we used the assumption that  $c' > e$  and  $\lambda$  is a large constant.  $\square$

**Lemma F.6** (Slight variation of Vershynin's theorem (Poisson instead of Bernoulli)). *Consider a random matrix  $A$  of size  $M \times M$ , where each entry follows an independent Poisson distribution  $A_{i,j} \sim \text{Poi}(P_{i,j})$ . Define  $d_{\max} = M \max_{i,j} P_{i,j}$ . For any  $r \geq 1$ , the following holds with probability at least  $1 - M^{-r}$ . Consider any subset consisting of at most  $10 \frac{M}{d_{\max}}$ , and decrease the entries in the rows and the columns corresponding to the indices in the subset in an arbitrary way. Then for some universal large constant  $c$  the modified matrix  $A'$  satisfies:*

$$\|A' - \mathbb{E}A\| \leq Cr^{3/2}(\sqrt{d_{\max}} + \sqrt{d'}),$$

where  $d'$  denote the maximal row sum in the modified random matrix.

*Proof.* The original proof in [33] is for independent Bernoulli entries  $A_{i,j} \sim \text{Ber}(P_{i,j})$ . The specific form of the distribution is only used when bounding the  $\ell_\infty \rightarrow \ell_1$  norm of the adjacency matrix by applying Bernstein inequality:

$$\Pr\left(\sum_{i,j=1}^M X_{i,j} > M^2 t\right) \leq \exp\left(\frac{M^2 t^2 / 2}{\frac{1}{M^2} \sum_{i,j}^M P_{i,j} + t/3}\right)$$

where  $X_{i,j} = (A_{i,j} - \mathbb{E}[A_{i,j}])x_i y_j$  for any fixed  $x_i, y_j \in \{+1, -1\}$ .

Recall that a random variable  $X$  is sub-exponential if there are non-negative parameters  $(\sigma, b)$  such that  $\mathbb{E}[e^{t(X - \mathbb{E}[X])}] \leq e^{t^2 \sigma^2 / 2}$  for all  $|t| < \frac{1}{b}$ . Note that a Poisson variables  $X \sim \text{Poi}(\lambda)$  has sub-exponential tail bound with parameters  $(\sigma = \sqrt{2\lambda}, b = 1)$ , since

$$\log(\mathbb{E}[e^{t(X - \lambda)}]) = (\lambda(e^t - 1) - \lambda t) - \lambda t^2 \leq 0, \text{ for } |t| < 1.$$

Therefore, when the entries are replaced by independent Poisson entries  $A_{i,j} \sim \text{Poi}(P_{i,j})$ , we can apply Bernstein inequality for sub-exponential random variables to obtain similar concentration bound:

$$\Pr\left(\sum_{i,j=1}^M X_{i,j} > M^2 t\right) \leq \exp\left(\frac{M^2 t^2 / 2}{\frac{1}{M^2} \sum_{i,j}^M \text{Var}(X_{i,j}) + bt}\right) \leq \exp\left(\frac{M^2 t^2 / 2}{2 \frac{1}{M^2} \sum_{i,j}^M P_{i,j} + t}\right).$$

The same arguments of the proof in [33] then go through.  $\square$